I Trattamenti Monetari non Pensionistici:

Approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10

Anni 1999-2001

Vincenzo Spinelli
Massimiliano Tancioni

Istat
DIST
DCSS-SIP

Sintesi: il documento è dedicato alla descrizione delle procedure di acquisizione ed elaborazione dell'informazione elementare sui Trattamenti Monetari di tipo non Pensionistico (TMNP) erogati dai datori di lavoro per conto dell'INPS (erogazione indiretta).
La presentazione ripercorre la sequenza delle operazioni effettuate a partire dall'acquisizione delle basi informative da fonte amministrativa alla loro riorganizzazione, passando attraverso le fasi di controllo ed imputazione degli importi mancanti ed il consolidamento rispetto ad alcune variabili di classificazione (settore di attività e dimensione occupazionale dell'impresa erogatrice).
La presentazione stabilisce pertanto l'insieme delle azioni che si sono ritenute necessarie e, allo stato attuale dell'informazione disponibile, sufficienti alla produzione del dato statistico finale.
La dimensione e la complessità dell'informazione contenuta negli archivi ha suggerito la realizzazione di un approccio computazionale il più possibile automatico, eliminando quindi la necessità di interventi "manuali"[1].
L'automazione dei processi di elaborazione permette inoltre una riduzione sostanziale dei tempi di produzione e di rilascio del prodotto finale.
Per ogni operazione descritta vengono specificate le considerazioni logiche giustificative, l'insieme delle funzioni (logiche o matematiche) utilizzate e il processo di implementazione informatica.

Parole chiavi*: Qualità e controllo della qualità nell'uso dei dati amministrativi, elaborazione dei dati.*

Abstract*: this work provides a description of the procedures of acquisition and conditioning of administrative elementary data on Non Pension Cash Benefits (NPCB) from the DM10 archive of INPS (Italy's National Social Security Institute). The description follows the sequence of the operations implemented, from the acquisition of the administrative sets to their conditioning with particular emphasis on the quality control. We describe the sequence of actions and controls that we deem as necessary and – on the basis of the available information - sufficient for the production of the final statistics.*
The high volume and complexity of data involved suggest the implementation of a fully automated computational approach, thus minimizing the necessity of "manual" actions on the archive. This approach allows a drastic reduction of the required production time.
For each step of the process, we specify the set of logical and mathematical considerations.

Keywords: *Quality and quality assurance in the use of administrative data, Data processing.*

[1] Per una rassegna di argomenti ed esperienze nel trattamento degli archivi amministrativi si veda Falorsi, Pallata e Russo 2003.

Indice

1. Premessa

Questo documento fornisce una dettagliata descrizione di alcune procedure automatiche implementate per il trattamento degli archivi amministrativi di dati elementari DM10 dell'INPS (Istituto Nazionale di Previdenza Sociale). Questi archivi contengono dati mensili sull'universo della spesa anticipata ai dipendenti del settore privato per i Trattamenti Monetari Non Pensionistici (TMNP). L'attenzione viene rivolta sui due aspetti centrali del processo elaborativo:

i. Identificazione e stima dei dati mancanti, con un approccio logico-longitudinale, relativi alla spesa a livello di posizione contributiva (matricola INPS);
ii. Identificazione e correzione non parametrica dei potenziali valori anomali nelle stime occupazionali a livello di impresa (numero di dipendenti medi nell'anno di riferimento).

Il primo aspetto riguarda la ricostruzione della serie temporale dei dati ed il loro consolidamento su base annuale. L'identificazione e l'imputazione dei valori mancanti viene effettuata tramite procedure automatiche sulla base sia di considerazioni logiche sia di informazioni longitudinali. Poiché gli eventi che danno origine alla spesa non sono stabili nel tempo, anche i dati connessi mostrano questa caratteristica. Ciò implica che procedure che si basano esclusivamente su informazioni relative alla dimensione temporale possono risultare non del tutto soddisfacenti. Infatti, esse non considerano elementi, oltre quelli temporali, che sono importanti da un punto di vista logico nella ricostruzione dei processi osservati. Per tener conto di questi potenziali inconvenienti, e assumendo che non siano disponibili informazioni da altre fonti informative, è stato definito un processo di individuazione dei dati mancanti basato su sequenze di vincoli logici operanti a monte di tecniche longitudinali. I dati che soddisfano i vincoli logici risultano essere un raffinamento dell'insieme di partenza. Le tecniche di identificazione e stima longitudinale dei dati mancanti vengono applicate solo su questo insieme ridotto.
Il secondo aspetto riguarda l'identificazione e la correzione dei potenziali valori anomali su dati occupazionali. L'uso di dati occupazionali è legato alla produzione di statistiche basate sulle classi dimensionali. È necessario osservare che le fonti dei dati occupazionali non sono direttamente collegabili al nostro insieme di riferimento, e quindi la comparabilità dei dati diventa un problema se il rapporto spesa/dimensione viene considerato come criterio di valutazione.
La procedura di individuazione dei valori anomali si basa su condizioni di ottimalità. Invece di considerare un approccio tradizionale (basato sulla definizione soggettiva di intervalli di confidenza), abbiamo sviluppato una procedura di selezione iterativa dei dati anomali basata su condizioni di ottimalità "quasi neutrali". La sostanziale neutralità della selezione si ottiene dall'utilizzo di criteri di valutazione per le condizioni di terminazione, la cui definizione deriva da un processo iterativo sugli stessi dati. Si ritiene che questo approccio riduca la componente soggettiva del processo di identificazione dei valori anomali, poiché l'aspetto determinante non è una ipotesi a priori sulla distribuzione sottostante, ovvero una misura per la massima dispersione dei dati.
Si assume che l'informazione TMNP sia esatta e completa e quindi costituisca un termine di confronto per il dato occupazionale; si assume inoltre che esiste un vincolo probabilistico tra la spesa TMNP e il numero di dipendenti e quindi

l'esistenza di una distribuzione definita nei i suoi momenti caratteristici. Queste ipotesi permettono di utilizzare la spesa media per dipendente come variabile di riferimento e, identificati i dati anomali, di imputarli attraverso il rapporto tra la spesa osservata e la spesa media per dipendente.

2. Breve descrizione degli archivi DM10

L'informazione di base è contenuta nei modelli di dichiarazione mensile DM10 compilati dalle imprese facenti pagamenti indiretti per conto dell'INPS[2].
Gli archivi DM10 vengono acquisiti in formato data-set SAS mensile. Essi contengono l'universo dell'informazione desumibile dai modelli di dichiarazione osservati *a 13 mesi dal periodo di riferimento*. Il vincolo temporale è stato stabilito sulla base di considerazioni sulla dinamica del popolamento degli archivi, descritte più in dettaglio in una sezione specifica.
I modelli DM10 si articolano in quattro quadri di dettaglio.
Il quadro A contiene le informazioni relative all'anagrafica di impresa, il quadro B contiene informazioni relative all'occupazione, il quadro C informazioni relative alle retribuzioni, il quadro D informazioni relative alle prestazioni anticipate dai datori di lavoro per conto dell'INPS.
La nostra attenzione si concentrerà esclusivamente sulle informazioni contenute nei quadri A e D (d'ora in avanti QA e QD), poiché essi costituiscono la base di partenza per l'applicazione del sistema di classificazione dei TMNP (SCPM)[3].

3. Classificazione: prestazioni, trattamenti e forme di pagamento

Prima di esporre le procedure di riorganizzazione del dato di partenza è necessario illustrare sinteticamente il sistema di classificazione utilizzato.
Il quadro D dei modelli mensili DM10 relativi al triennio 1999-2001 contiene 540 codici identificativi di forme di pagamento, di cui solo alcune sono configurabili come TMNP[4]. Sulla base della classificazione SCPM, vengono selezionate 56 forme di pagamento, riferibili a 13 trattamenti distinti, a loro volta riconducibili alle seguenti 10 tipologie di prestazione:

→ *Indennità di malattia generica*
→ *Indennità di malattia specifica (TBC)*
→ *Indennità di maternità (obbligatoria e facoltativa, per allattamento)*
→ *Indennità per handicap (famiglia)*
→ *Indennità per handicap (lavoratore)*
→ *Assegni al nucleo familiare (ANF)*
→ *Assegni al nucleo familiare vecchio tipo (ANF-VT)*
→ *Assegni per congedo matrimoniale*
→ *Integrazioni salariali (CIGO – CIGS)*
→ *Piani di inserimento professionale (PIP)*

[2] Per la descrizione di altre esperienze di utilizzo degli archivi DM10 si veda Baldi, Cimino, Pallata, Succi e Tuzi, 2001.
[3] Si veda Consolini e De Carli, 2002 (Appendice 1).
[4] Tali forme di pagamento traggono origine dalla normativa e da provvedimenti amministrativi aventi spesso carattere temporaneo.

La determinazione della struttura (prestazione-trattamenti-forme di pagamento) è stata effettuata, in prima approssimazione, utilizzando la definizione di TMNP[5], come descritto nella tavola 3.1.
La struttura relazionale è stata inoltre sottoposta ad un processo di verifica di robustezza eseguita sulle forme di pagamento, cioè al massimo livello di disaggregazione desumibile dagli archivi. Il codice relativo alla singola forma di pagamento è giudicato *ammissibile* sulla base della soddisfazione delle seguenti condizioni:

→ *appartenenza* ad una delle tipologie stabilite, date le definizioni;
→ *esaustività* (robustezza all'estensione concettuale) e non duplicazione;
→ *vigenza del provvedimento* dante origine nel periodo di riferimento 1999-2001.

Vengono inoltre ritenute ammissibili quelle forme di pagamento per le quali si sia registrata, data la soddisfazione delle prime due condizioni, l'esistenza di osservazioni indipendentemente dalla vigenza del provvedimento dante origine, ossia indipendentemente dalla soddisfazione della terza condizione[6].
L'elaborazione di dati relativi a periodi successivi a quelli considerati in questa sperimentazione dovrà valutare la possibilità che nuovi interventi legislativi ne abbiano modificato lo schema di classificazione.

4. Condizionamento dell'informazione di partenza

Sulla base della classificazione esposta sopra, i dati provenienti dagli archivi DM10 vengono riorganizzati in due data-set relativi alle informazioni in QA e QD utili all'indagine. Tale operazione viene effettuata attraverso l'esecuzione di due procedure distinte: *QA.sh* e *QD.sh.*

4.1. Creazione dell'anagrafica di impresa

La procedura *QA.sh* filtra l'informazione di partenza di livello anagrafico su base mensile producendo l'archivio annuale delle imprese facenti dichiarazioni DM10.
Dal momento che le unità statistiche del quadro D vengono individuate dal legame che esse hanno con l'informazione presente nel quadro A, il numero delle imprese contenute nell'archivio anagrafico è non inferiore al numero di quelle presenti nel quadro relativo ai trattamenti anticipati per conto dell'INPS.
L'anagrafica di impresa definisce l'universo di riferimento. L'informazione di livello anagrafico, organizzata per anno, mese di erogazione e codice identificativo INPS, o matricola[7], riporta i dati sulle seguenti variabili:

→ codice fiscale o partita IVA;
→ ragione sociale;
→ forma giuridica;

[5] Per essa, si confronti Consolini (2000) e Consolini e DeCarli (2001).
[6] Questa eccezione viene adottata ai fini della considerazione di trattamenti per arretrati relativi a codici teoricamente inattivi, per i quali si registra le presenza di erogazioni nel periodo di riferimento.
[7] La matricola identifica una posizione contributiva dell'impresa verso l'INPS.

Tavola 3.1. Denominazione e raccordo tra forme di pagamento, trattamenti, prestazioni e funzione SESPROS[8]

Funzione	codice	Prestazione	codice	Trattamento	codice	Forme di pagamento
MALATTIA	*010*	*MALATTIA GENERICA*	01	Malattia generica	0052	indennità economiche erogate mese rif. denuncia
					E778	differenza indennità di malattia
	040	*MALATTIA SPECIALE TBC*	02	Malattia speciale TBC	0054	indennità economiche erogate mese rif. denuncia
FAMIGLIA	*141*	*MATERNITA'*	03	Maternità obbligatoria	0053	importi erogati astensione obbl mese rif la denuncia
					E779	differenza indennità di maternità obbligatoria
			04	Maternità facoltativa	L050	indennità di maternità facoltativa
					L055	differenza indennità di maternità facoltativa
			05	Maternità allattamento	D800	indennità riposi per allattamento
					D900	residuo credito dell'indennità
	142	*HANDICAP FAMIGLIA*	06	Handicap	L053	indennità fac agevolazioni per genitori minori hand
					L054	indennità fac permessi giornalieri per genitori minori hand
					L056	indennità fac permessi tre giorni mensili genitori/parenti sogg hand
					L070***	indennità congedo str genitori soggetti hand gravi accertati 5 anni
	160	*ASS.NUCLEO FAMILIARE*	07	Ass. Nucleo Familiare	0035	erogazioni correnti
					F240	differenza ANF caratisti e armatori
					G821***	ANF di cui CIGO
					H301	ANF indennità TBC
					L036	arretrati ANF
					T131	assegni familiari su indennità
					T133	assegni familiari su indennità arretrati
					T135	maggiorazione assegni familiari
					(T151*)	ANF portuali
					(T152)	ANF portuali arretrati
	170	*ASSEGNI FAMILIARI V. TIPO*	08	Assegni familiari V.T.	H017	maggiorazioni assegni familiari
					H300*	assegni familiari caratisti e armatori
	200	*CONGEDO MATRIMOLIALE*	09	Congedo matrimoniale	L051	assegno congedo matrimoniale
					L052	differenza assegno congedo matrimoniale
INVALIDITA'	*300*	*HANDICAP LAVORATORE*	10	Handicap	L057	indennità fac permessi due ore giornaliere lavoratori handicap
					L058	indennità fac permessi tre giorni mensili lavoratori handicap
DISOCCUPAZIONE	*221*	*INTEGRAZIONI SALARIALI*	11	Integrazioni CIGO	0039	erogazioni CIGO
					E200	ratei CIGO non soggetti a CTR addizionale
					E800	ratei CIGO soggetti a CTR addizionale
					G400	integrazioni salariali ordinarie non soggette a CTR addizionale
					G820***	CIGO G8
					V880	CIGO arretrati
			12	Integrazioni CIGS	0040	erogazioni CIGS
					F500	ratei CIGS non soggetti a CTR addizionale
					G600	integrazioni salariali straordinarie non soggette a CTR addizionale
					G602****	maggiorazioni integrazioni salariali straordinarie
					G603	integrazioni salariali straordinarie e contratti di solidarietà
					G604	maggiorazione 10% integrazioni salariali straord e contr solidarietà
					G605*****	integrazioni salariali straord ridotte 10% in proroga
					G606*	integrazioni salariali straord ridotte 10% in proroga
					G607*	integrazioni salariali straord ridotte 10% in proroga
					G608*	integrazioni salariali straord ridotte imprese in amm.ne straord
					G609*	integrazioni salariali straord ridotte consorzi agrari
					T145	integrazioni salariali straordinarie EFIM
					V890	CIGS arretrati CCNL
	281	*PIANI INS. PROFESSIONALE*	13	Piani Ins. Professionale	R770	indennità base pip
					R771**	indennità base pip orario concentrato
					R772**	indennità base pip giovani residenti sicilia (e i calabri poveretti?)
					R773**	indennità base pip orario concentrato giovani resid regione Sicilia
					R780	indennità aggiuntiva pip
					R781**	indennità aggiuntiva pip orario concentrato
					R782**	indennità aggiuntiva pip giovani residenti regione Sicilia
					R783**	indennità aggiuntiva pip orario concentrato giovani resid reg Sicilia

Nota: * = codici ammissibili solo su base 1999; ** = codici ammissibili solo dal 2000; *** = codici ammissibili solo dal 2001; **** = codici ammissibili fino al 2000; ***** = codici ammissibili per effetto proroga di durata sconosciuta; (Codice forma di pagamento) = codici ammessi in virtù di presenza osservazioni anche se teoricamente inattivi.

[8] Si veda Eurostat (1996).

→ codice statistico contributivo;
→ codice identificativo del settore di attività economica secondo la definizione INPS;
→ data di costituzione dell'azienda;
→ sede INPS di riferimento;
→ codice ATECO[9] identificativo del settore di attività economica;

La sperimentazione sui dati riferiti agli anni 1999, 2000 e 2001 ha rilevato, rispettivamente, la presenza di 1.410.673, 1.460.133 e 1.500.720 matricole INPS distinte.

4.2. Creazione dell'archivio dei dati di spesa TMNP

La procedura *QD.sh* filtra l'informazione di partenza relativa alla spesa TMNP su base mensile, producendo l'archivio annuale delle imprese che effettuano dichiarazioni di pagamento TMNP all'INPS. Il numero delle imprese contenute nell'archivio QD è in tal caso *non superiore* al numero di quelle presenti nell'archivio QA. L'informazione di spesa, organizzata per anno, mese e codice identificativo INPS, riporta i dati sulle seguenti variabili[10]:

→ funzione SESPROS
→ prestazione
→ trattamento
→ forma di pagamento
→ importo

La sperimentazione su dati riferiti agli anni 1999, 2000 e 2001 ha riscontrato, rispettivamente, la presenza di 814.581, 831.940 e 834.235 matricole INPS distinte. Tale numerosità è riferita alle imprese per le quali si riscontra la presenza di almeno una forma di pagamento in almeno un mese dell'anno di riferimento.

L'insieme delle operazioni descritte ed effettuabili attraverso l'esecuzione delle procedure *QA.sh* e *QD.sh* avviene su base mensile, in corrispondenza dell'acquisizione dei singoli archivi DM10. Quando è pervenuta l'informazione di base per tutti i dodici mesi di un anno solare, essa viene raccolta in un archivio annuale di anagrafica e in uno di dettaglio relativo agli importi associati alle singole forme di pagamento.

Quest'ultimo costituisce *l'universo dei microdati utili* ai nostri fini, ai quali non è stata ancora imposta alcuna procedura di verifica e correzione sugli importi.

4.3. Ricostruzione dell'impresa e attribuzione del codice identificativo del settore di attività economica

La matricola INPS identifica la specifica posizione contributiva piuttosto che la singola impresa. Ciò rende necessaria, ai fini della definizione di un archivio avente unità di analisi compatibili con quelle di interesse statistico, la traduzione dell'archivio INPS (che è un archivio in unità contributive) in un archivio in unità di impresa. La ricostruzione delle corrispondenze (potenzialmente

[9] Si veda Istat (1991).
[10] Le voci indicate fanno riferimento alla Tavola 3.1.

multiple) matricola-impresa viene effettuata dal servizio ARC dell'Istat attraverso procedure standard proprie.
Il codice identificativo del settore di attività economica attribuito sulla base dell'informazione originaria DM10 viene mantenuto esclusivamente al fine di coprire le potenziali cadute informative associate al processo di ricostruzione dell'impresa a muovere dalla matricola INPS.
L'identificazione dell'impresa implica l'attribuzione del codice identificativo dell'attività economica come da archivio delle imprese attive (ASIA). Il processo di ricostruzione dell'impresa costituisce un elemento di validazione dell'archivio soprattutto nel senso che lo rende compatibile con le unità di analisi e le definizioni Istat.
La sperimentazione per gli anni qui considerati ha tuttavia evidenziato che il processo di ricostruzione non è esente da cadute informative. Sebbene in linea di principio sia lecito ipotizzare che l'universo delle imprese facenti riferimento alle matricole presenti negli archivi DM10 – a meno di quelle relative al settore agricolo - sia un sottoinsieme dell'universo delle imprese contenute nell'archivio ASIA, si è verificato che esiste una certa numerosità di posizioni contributive per le quali il servizio ARC non riesce ad effettuare l'identificazione e la ricostruzione dell'impresa.
Il problema, nei tre anni considerati, ha riguardato, rispettivamente, 10.782 (1,3%), 14.189 (1,7%) e 10.494 (1,3%) matricole INPS.
Le matricole per le quali non si dispone dell'aggancio con l'archivio ASIA vengono considerate imprese. Per esse si assume pertanto che esista una corrispondenza univoca tra posizione contributiva e impresa, col risultato che l'impresa resta identificata dalla singola matricola INPS.
Il codice identificativo del settore di attività economica, relativamente a queste imprese, viene dedotto dall'informazione contenuta negli archivi DM10. Ciò ha reso necessaria la realizzazione dei programmi di lettura e traduzione dei codici identificativi del settore di attività economica.
I dati relativi al codice ATECO[11] vengono in una prima fase attribuiti attraverso l'utilizzo di una procedura automatica di controllo e correzione del codice identificativo del settore di attività economica secondo la definizione Istat. In assenza di informazione sul codice identificativo Istat esso viene ottenuto, dove ne esista informazione, dalla traduzione del codice statistico dell'INPS in classificazione ATECO81 (la compatibilità fa riferimento a questa classificazione) e successivamente nella classificazione ATECO91.
È stato altresì riscontrato che, dato un anno di riferimento, il settore di attività economica riferibile ad una data posizione contributiva INPS può cambiare nello scorrere dei mesi. Poiché le unità di impresa contenute nell'archivio ASIA si riferiscono alla dinamica demografica annuale delle imprese, il dato di ATECO mensile deve essere consolidato allo stesso livello annuale.
Ciò viene eseguito da una procedura automatica che sfrutta l'ipotesi per la quale il codice di attività economica prevalente a livello annuale è quello relativo al mese più recente dell'anno preso in considerazione, nel quale il codice ATECO è non nullo. Si utilizza pertanto un criterio convenzionale che privilegia l'informazione più recente.

[11] Istat, 1991.

Nonostante le procedure implementate, non sempre è possibile la ricostruzione dell'attività economica dell'impresa. Per gli anni 1999, 2000 e 2001, le imprese non ripartibili rispetto all'ATECO sono, rispettivamente, 65, 1054 e 19[12].
L'appartenenza dell'impresa ad uno dei settori di attività economica ATECO, a meno delle cadute informative, è quindi compatibile con la ricostruzione delle corrispondenze matricole-impresa effettuata dal servizio ARC.

4.4. Definizione della classe dimensionale delle imprese ricostruite
Lo stesso vincolo di compatibilità deve essere rispettato, dove possibile, anche per il dato occupazionale, in base al quale si definiscono le 6 classi dimensionali utilizzate nell'analisi, come definite nella tavola 4.1. Le imprese ricostruite sulla base dell'archivio ASIA ereditano il dato occupazionale definito in termini di numero medio annuale di lavoratori dipendenti.

Tavola 4.1. Definizione delle classi dimensionali

Numero medio dipendenti	Classe dimensionale
$[0,6[$	*1*
$[6,10[$	*2*
$[10,20[$	*3*
$[20,100[$	*4*
$[100,250[$	*5*
250 *e oltre*	*6*

Nel caso di mancato aggancio tra matricole INPS ed impresa ASIA, analogamente al caso ATECO, il dato occupazionale viene stimato direttamente dalle informazioni contenute negli archivi DM10. Questi archivi sono mensili, e la dimensione media di un'impresa, valida nell'anno di riferimento, viene ottenuta come media aritmetica dei valori occupazionali mensili[13].

5. Procedure di controllo di qualità e imputazione dei dati elementari di spesa mancanti

Nella fase che segue alla ricostruzione dell'impresa si procede al controllo della qualità dei microdati per singolo codice di trattamento e alla loro ricostruzione. Ciò corrisponde all'implementazione di due procedure congiunte, la prima dedicata alla identificazione dei dati mancanti su base mensile, la seconda alla loro imputazione.
La natura dei TMNP non permette una analisi agevole, poiché essi originano dal verificarsi di eventi per i quali non si dispone di informazioni ulteriori esterne all'archivio. Inoltre, i trattamenti hanno, nella maggioranza delle tipologie, carattere temporaneo, il che compromette l'applicabilità di metodi longitudinali

[12] Il picco di numerosità nelle imprese non ripartibili per ATECO osservato per l'anno 2000 è dovuto a problemi emersi nel processo di ricostruzione dell'impresa. La veste sperimentale di questa analisi ha suggerito di trascurare questo problema, che verrà risolto nelle indagini a regime.
[13] Anche se in alcuni mesi il dato dimensionale non è presente, la media aritmetica viene riferita comunque a 12 mesi.

estrapolativi, basati cioè sulla "memoria" (deterministica o stocastica) dei processi.
Si assume tuttavia che per alcuni casi sia ragionevole ipotizzare una relativa stabilità mensile dei trattamenti, ritenendo permanente (almeno nell'ottica di breve periodo qui adottata[14]) la causa dante origine allo specifico TMNP.
Come criterio di prima selezione (operante al livello di singolo trattamento), si assume che il controllo sia praticabile in relazione a quei regimi per i quali la causa dante origine al trattamento, quindi la spesa, abbia carattere permanente.
Sotto questa ipotesi, si ritiene che i trattamenti "Assegno al Nucleo Familiare" (07), "Handicap famiglia" (06) e "Handicap lavoratore" (10) possano essere sottoposti all'analisi di tipo longitudinale, applicabile a livello di singola posizione contributiva e singola forma di pagamento.
Fissato l'intervallo temporale di riferimento, si assume inoltre che l'informazione in esso presente sia sufficiente all'individuazione dei potenziali dati mancanti.
Viene impostata una procedura automatica di correzione basata su una tecnica mista logico-longitudinale che sfrutta tutta l'informazione interna all'archivio per l'anno di riferimento. La metodologia utilizzata può essere rappresentata in due fasi.
La *prima* è di livello logico e produce l'individuazione dei dati potenzialmente mancanti, attraverso la soddisfazione di batterie di vincoli di confrontabilità degli importi relativi a mesi diversi e/o forme di pagamento diverse relative allo stesso trattamento.
La *seconda* procedura, che si applica dopo l'individuazione dei casi per i quali si assume la potenziale esistenza di dati mancanti (data la soddisfazione dei vincoli precedenti), procede all'imputazione del dato sulla base dell'informazione relativa alla stessa impresa nei mesi per i quali essa è presente (la "donazione" avviene pertanto nel dominio del tempo e non in quello sezionale). In entrambe le procedure (individuazione e correzione) vengono applicati criteri di tolleranza variabile che permettono la calibrazione dell'intervento.
Con riferimento al trattamento 07, i codici di forme di pagamento sottoposti all'analisi sono: 0035-L036, T131-T133, T151-T152 (cnfr. Tavola 3.1).
Con riferimento al trattamento 06, i codici di forme di pagamento sottoposti all'analisi sono: L053, L054, L056, L070 (cnfr. Tavola 3.1).
Con riferimento al trattamento 10, i codici di forme di pagamento sottoposti all'analisi sono: L057, L058, (cnfr. Tavola 3.1).
Di seguito sono descritte le procedure di identificazione ed imputazione per i codici afferenti ai trattamenti 07, 06 e 10.
Per una migliore comprensione delle procedure implementate, si consideri che l'analisi di livello logico definisce l'insieme *potenziale* dei mesi contenenti dati mancanti, mentre l'analisi quantitativa definisce l'insieme di intervento, cioè l'insieme dei mesi per i quali, sulla base di specifiche ipotesi di confrontabilità, si assume l'esistenza di dati mancanti.

5.1. Procedura di controllo e imputazione dei dati del trattamento "07"

Ai fini dell'implementazione delle procedure descritte di seguito, l'informazione ottenuta a completamento delle elaborazioni della procedura precedente viene riorganizzata in una struttura matriciale avente un numero di righe pari al

[14] L'approccio di breve periodo è reso evidente dal fatto che le operazioni di identificazione ed imputazione dei dati mancanti vengono effettuate sui singoli dati mensili.

numero di matricole INPS ed un numero di colonne pari al numero di mesi considerati[15]. All'interno delle celle così definite viene attribuito un codice dicotomico per l'individuazione dei dati potenzialmente mancanti (1 = esiste trattamento, 2 = non esiste trattamento).

5.1.1 Definizione dell'insieme potenziale di imputazione

Sia $T = \{1,2,\ldots,M\}$ l'intervallo discreto del tempo relativo ai mesi *m* degli anni considerati, non necessariamente consecutivi. Per ogni impresa e per ogni trattamento, la definizione dell'insieme dei mesi *potenzialmente* soggetti ad imputazione viene derivato dal soddisfacimento delle seguenti condizioni:

→ l'informazione relativa ad un generico mese $m \in T$ è potenzialmente mancante se e solo se:

a) non è stato effettuato un pagamento diretto (valore 0)
b) $\exists\, n \in T : n < m$, per cui sia stato effettuato un pagamento diretto (valore 1)
c) $\exists\, \nu \in T : \nu > m$, per cui sia stato effettuato un pagamento diretto (valore 1)

La tabella che segue fornisce alcuni esempi delle fattispecie considerate nei vincoli di inclusione nell'insieme potenziale, dove si ipotizza la presenza di informazioni relative ad un solo anno, quindi 12 mesi[16].

Tavola 5.1. Alcuni esempi di fattispecie di inclusione (casi 7-12) ed esclusione nell'insieme potenziale (casi 1-6)

	mesi											
caso	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	1	1	1	1	1	1	1	1	1
2	***0***	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	**0**
4	**0**	1	1	1	1	1	1	1	1	1	1	**0**
5	**0**	**0**	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	**0**	**0**
7	1	**0**	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	**0**	1
9	1	**0**	**0**	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	**0**	**0**	1
11	**0**	1	**0**	1	1	1	1	1	1	1	1	1
12	**0**	1	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

[15] Si consideri che le procedure di controllo e imputazione descritte in questo paragrafo si riferiscono alle matricole presenti nell'archivio QD con scarico a 13 mesi dal mese di riferimento. Sperimentazioni INPS sugli archivi riferiti agli anni 1997 e 1998 hanno stimato che esse costituiscono circa il 95% del totale delle matricole effettivamente compilanti il quadro D.

[16] Le procedure attualmente implementate applicano l'approccio descritto su intervalli temporali potenzialmente più ampi (fino a 10 anni).

Come si nota, i casi 2-6 non vengono inclusi nell'insieme potenziale poiché i mesi in cui non sono presenti pagamenti diretti ("non esistenze": valore 0) non trovano campi valorizzati passati o futuri nell'insieme di riferimento *T* ("esistenze": valore 1).
In particolare, le fattispecie 2 e 5 non soddisfano la condizione "b", le fattispecie 3 e 5 la condizione "c" e la fattispecie 4 entrambe.
Le fattispecie 11 e 12 vengono incluse nell'insieme potenziale ma solo con riferimento ai mesi mancanti non posti agli estremi dell'intervallo di riferimento, cioè al mese 3.
Questo primo insieme di restrizioni definisce una riduzione dell'insieme di riferimento. Il loro significato implicito è che si ritengono logicamente ammissibili (quindi esclusi dall'insieme), in assenza di altre informazioni, i dati mancanti agli estremi dell'intervallo, potendo essi essere spiegati dal presentarsi, o dal mancare, della causa dante origine al trattamento, mentre si ritengono potenzialmente inammissibili (quindi inclusi) i dati mancanti interni all'intervallo e pertanto compresi tra due "esistenze".
In effetti, sarebbe possibile sottoporre a verifica l'ammissibilità di "non esistenze" agli estremi intervallo effettuando una estensione dell'intervallo stesso a destra e a sinistra, cioè richiamando altra informazione relativa ad anni passati e futuri rispetto a quello di riferimento. Questa possibilità deve tuttavia essere valutata in relazione ai vantaggi e agli svantaggi ad essa associati.
Sebbene l'estensione del potenziale informativo dovrebbe produrre variazioni non negative della qualità del dato, essa induce altresì una maggiore complessità dell'analisi ed inoltre espone ai rischi di decadimento del contenuto informativo connessi alla distanza temporale su cui viene effettuato il confronto.
La valutazione degli svantaggi associati all'estensione assume rilevanza cruciale. In primo luogo, una maggiore complessità si associa ad una maggiore probabilità di incorrere in errori (questo ragionamento è alla base dei criteri informativi di scelta parsimoniosa, che necessariamente utilizzano un criterio soggettivo nella definizione dei pesi negativi attribuiti alla complessità); in secondo luogo, l'estensione dell'informazione verso destra (al futuro rispetto all'anno di riferimento) produce banalmente una analoga estensione dei tempi di produzione del dato. È evidente che questa seconda considerazione non si applica in relazione all'estensione a sinistra (al passato), nel qual caso si incontrerebbero esclusivamente gli svantaggi connessi alla maggiore complessità di applicazione della procedura.
A fini esemplificativi, si decide per questa applicazione la *non estensione* del set informativo, assumendo come *sufficiente*, l'informazione relativa al singolo anno di riferimento.

5.1.2. Definizione dell'insieme di intervento

Per ogni impresa e per ogni forma di pagamento, per la quale il soddisfacimento dell'insieme di condizioni di cui sopra abbia definito l'esistenza di mesi *potenzialmente* soggetti ad imputazione, l'identificazione dell'insieme di mesi per i quali *è ragionevole ritenere l'esistenza* di dati mancanti (cioè la "non esistenza" del dato) è definita dalla soddisfazione dei seguenti due ulteriori insiemi di condizioni (*a* e *b*):

caso sub a:

→ dati i mesi *potenzialmente* soggetti ad imputazione di dati mancanti per essi non si registrano trattamenti arretrati (L036), cioè:

$$\nexists\, h \in T : h > m, \text{ per cui sia stato effettuato un pagamento arretrato}$$

→ dato il soddisfacimento della condizione precedente, gli importi *i* associati ai mesi *n* e *v* devono essere di un ordine di grandezza "simile" (scostamento relativo < *k*), cioè:

$$\left|(i_v - i_n)/i_n\right| \le k$$

La tabella che segue fornisce alcuni esempi di fattispecie di inclusione nell'insieme di imprese per le quali si assume l'esistenza di dati mancanti sub caso *a*, compilata assumendo un valore di discriminazione $k = 0.05$ [17].

Tavola 5.2. Alcuni esempi di fattispecie di inclusione ed esclusione nell'insieme di correzione sub caso *a*

		mesi													
caso	fasi	1	2	3	4	5	6	7	8	9	10	11	12	deviazione (%)	inclusione
1	trattamento	1	0	1	1	1	1	1	1	1	1	1	1	-	-
1	arretrato	0	0	1	1	0	0	0	0	0	0	0	0	-	N
2	trattamento	1	1	1	1	1	1	1	1	1	1	0	1	-	-
2	arretrato	0	0	0	1	0	0	0	0	0	0	0	0	-	-
2	importo tratt.	100	100	100	100	100	100	100	100	100	100	0	120	20	N
3	trattamento	1	0	0	1	1	1	1	1	1	1	1	1	-	-
3	arretrato	1	0	0	0	0	0	0	0	0	0	0	0	-	-
3	importo tratt.	100	0	0	100	100	100	100	100	100	100	100	100	0	Y
4	trattamento	0	1	0	1	1	1	1	1	1	1	1	1	-	-
4	arretrato	0	0	0	0	0	0	0	0	0	0	0	0	-	-
4	importo tratt.	0	100	0	101	101	101	101	101	101	101	101	101	1	Y

Il caso 1 non soddisfa la prima condizione sub *a* poiché esiste un trattamento arretrato in un mese successivo a quello in cui si registra un valore nullo nel trattamento corrente. Il caso 2 non viene incluso nell'insieme di correzione poiché, pur essendo soddisfatta la prima condizione, non lo è la seconda, essendo lo scostamento assoluto degli importi passati e futuri rispetto al valore nullo maggiore del valore di tolleranza. Le fattispecie 3 e 4 vengono incluse nell'insieme di correzione poiché, essendo in entrambi i casi soddisfatta la prima condizione di assenza di trattamenti arretrati successivi al valore nullo (o ai valori nulli), i corrispondenti importi passati e futuri sono di un ordine di grandezza giudicato compatibile sulla base del valore di discriminazione adottato.

[17] Il processo di identificazione dei dati mancanti deve fornire un ragionevole compromesso tra la sensibilità del criterio di non esistenza del dato (le sequenze temporali devono essere il più possibile complete) ed il sostanziale mantenimento della spesa di partenza (non si vuole cambiare i totali di spesa in modo significativo). Il valore $k = 0.05$ è generalmente accettato nelle procedure di valutazione delle ipotesi. A seguito delle prove effettuate, questo valore del parametro k è quello che ha mostrato il miglior compromesso tra i due aspetti precedenti.

caso sub b:

→ dati i mesi *potenzialmente* soggetti ad imputazione per essi si registrano trattamenti arretrati (L036) in mesi *h* successivi, cioè:

$\exists h \in T : h > m$ per cui sia stato effettuato un pagamento arretrato;

→ dato il soddisfacimento della condizione precedente gli importi *i* dei pagamenti arretrati *h* non sono sufficienti alla copertura del valore da imputare, definito per interpolazione lineare nei trattamenti mensili diretti:

$$\sum_{h=m+1}^{M} i_h < \hat{i}_m \ , \ \hat{i}_m = i_v + \left[\left(i_n - i_v \right) \frac{m-v}{n-v} \right]$$

Questo secondo insieme di vincoli stabilisce in sostanza che l'informazione connessa ai trattamenti per arretrati (nel caso sub *a* l'*assenza* di arretrati nel periodo successivo alla "non esistenza", nel caso sub *b* l'incommensurabilità tra valori mancanti interpolati e valori dei trattamenti arretrati) e l'informazione sull'ordine di grandezza degli importi esistenti vengono utilizzati come ulteriori dispositivi di identificazione delle non esistenze. Evidentemente si assume che l'informazione per trattamenti arretrati sia vera e completa. Questa ipotesi è necessaria, dal momento che per essi non si dispone di elementi ulteriori di validazione, e sufficiente, poiché si è assunta l'esaustività, ai fini dell'indagine, dell'informazione relativa al solo periodo di osservazione.

5.1.3. Procedure di imputazione dei dati mancanti
Effettuata l'identificazione dell'insieme di mesi per i quali si assume l'esistenza di dati mancanti l'imputazione dei valori avviene nel modo seguente:

caso sub a:

→ ad ogni valore nullo si attribuisce un valore pari a quello linearmente interpolante i valori esterni alla/e non esistenza/e:

$$\hat{i}_m = i_v + \left[\left(i_n - i_v \right) \frac{m-v}{n-v} \right];$$

caso sub b:

→ l'imputazione del valore mensile mancante è derivata dalla soluzione del seguente problema di minimo[18]:

$$\min_{h=m+1,\ldots,M} \left\{ \sum_{j=m+1}^{h} i_j \geq \hat{i}_m \right\}$$

[18] Il problema descritto cerca di individuare il più piccolo intervallo temporale nel futuro ($\left[m+1, h \right]$) che presenta importi arretrati sufficienti a coprire il valore da imputare. Per una trattazione dei problemi di programmazione dinamica si veda Douglas Faires e Burden, (1998).

Le soluzioni possibili sono:

$$\begin{cases} h^* = M, \quad \hat{i}_m = \sum_{\mathrm{j}=m+1}^{\mathrm{h}^*=\mathrm{M}} i_j \quad \Rightarrow \quad i_m = \hat{i}_m, \; i_{h=m+1,\ldots,M} = 0 \\ h^* < M, \quad \hat{i}_m \le \sum_{\mathrm{j}=m+1}^{\mathrm{M}} i_j \quad \Rightarrow \quad i_m = \hat{i}_m, \; i_{h=m+1,\ldots,h^*-1} = 0, \; i_{h^*} = i_{h^*} - \left(i_m - \sum_{\mathrm{j}=\mathrm{m}+1}^{\mathrm{h}^*\text{-}1} i_j \right) \\ \nexists\, h^* \quad \Rightarrow \quad \hat{i}_m > \sum_{\mathrm{j}=m+1}^{\mathrm{M}} i_j \quad \Rightarrow \quad i_m = \hat{i}_m^{\hbar}, \; i_{h=m+1,\ldots,M} = 0 \end{cases}$$

dove

$$\hat{i}_m^{\hbar} = i_v + \left[\left(i_n - i_v \right) \frac{m-v}{n-v} \right], \text{ se } \left| (i_v - i_n) / i_n \right| \le r \,, \; r > k \,.$$

Quanto scritto sopra stabilisce in sostanza due fattispecie separate.
Nella prima esiste la soluzione al problema di programmazione dinamica descritto. Si ha che gli importi arretrati esistenti sono sufficienti a coprire gli importi mensili diretti stimati. In tal caso gli importi stimati vengono imputati e allo stesso tempo gli importi arretrati vengono tutti azzerati tranne l'ultimo utile alla copertura, che viene ridotto dell'importo residuo da imputare. È evidente che, nel caso in cui l'importo residuo equivale all'ultimo importo utile, tutta la successione di arretrati individuati dall'indice ottimo vengono azzerati. Le correzioni sui mesi vengono pertanto effettuate in successione.
Si noti che in tal caso non si produce una alterazione di spesa a livello di trattamento, poiché si è esclusivamente eseguita una redistribuzione tra forme di pagamento. Si assume pertanto che, nel caso in cui la spesa per arretrati presente nell'intervallo di riferimento sia complessivamente sufficiente a coprire la spesa diretta stimata associata ai dati mancanti, la correzione non debba alterare la spesa complessiva, ma solo la distribuzione per forme di pagamento.
Sebbene tale alterazione distorca l'immagine reale delle forme di pagamento, essa viene comunque eseguita in vista del suo significato di validazione dell'archivio, ed in considerazione del fatto che il dettaglio minimo a cui siamo interessati è il trattamento e non la forma di pagamento.
Nella seconda fattispecie il problema di programmazione non ha una soluzione, poiché l'indice temporale ottimo non appartiene all'insieme di riferimento. In altri termini, gli importi arretrati esistenti non sono sufficienti a coprire gli importi mensili diretti stimati. In tal caso l'imputazione viene eseguita per interpolazione rilassando il vincolo di confrontabilità e gli importi arretrati vengono tutti azzerati. La variazione di spesa che si produce, al livello di trattamento, è data dalla differenza, sempre positiva, tra la somma degli importi interpolati imputati e la somma degli importi arretrati successivi alla prima "non esistenza" calcolata prima del loro azzeramento.
La procedura di imputazione sostanziale, analoga al caso sub *a*, viene attivata solo in caso di mancata soluzione del problema di minimo. Esso è pertanto un ulteriore criterio di valutazione di eseguibilità dell'imputazione che utilizza l'informazione relativa alla spesa arretrata. Come si rileva dalle definizioni formali, il vincolo di confrontabilità posto a condizione di esecuzione dell'interpolazione viene in tal caso rilassato ($r > k$). Il rilassamento del vincolo, come del resto la sua definizione, è un elemento soggettivo la cui opportunità

deriva dalla considerazione del fatto che il confronto si effettua potenzialmente su importi già interpolati.
Dato l'ultimo vincolo stabilito dalla seconda batteria di assunzioni, il valore (o i valori) imputato non potrà discostarsi, in valore assoluto, in misura maggiore del valore di tolleranza dai valori adiacenti preesistenti, che pertanto ne definiscono l'intervallo dei valori di ammissibilità dell'imputazione.
La tabella che segue riporta le imputazioni effettuate per le fattispecie esemplari di inclusione di cui alla tavola 5.2 sotto caso *a*:

Tavola 5.3. Imputazione di valori mancanti per le fattispecie esemplari di cui alla tavola 5.2

		mesi												
caso	fasi	1	2	3	4	5	6	7	8	9	10	11	12	deviazione (%)
1	importo tratt.	100	0	0	100	100	100	100	100	100	100	100	100	0
1	imputazione	0	100	100	0	0	0	0	0	0	0	0	0	-
1	nuovo imp.	100	100	100	100	100	100	100	100	100	100	100	100	0
2	importo tratt.	0	100	0	101	101	101	101	101	101	101	101	101	1
2	imputazione	0	0	100,5	0	0	0	0	0	0	0	0	0	-
2	nuovo imp.	0	100	100,5	101	101	101	101	101	101	101	101	101	0,5

5.2. Procedure di controllo e correzione dei codici delle tipologie di trattamento "06 e 10"

5.2.1. Definizione dell'insieme potenziale
Analogamente al caso precedente, per ogni impresa e per ogni forma di pagamento si definisce, attraverso una prima batteria di vincoli, l'insieme di mesi che si ritengono *potenzialmente* soggetti a dati mancanti:

→ un mese $m \in T$ è potenzialmente imputabile se e solo se:

a) non è stato effettuato un pagamento diretto (valore 0)
b) $\exists\, n \in T : n < m$, per cui sia stato effettuato un pagamento diretto (valore 1)
c) $\exists\ \nu \in T : \nu > m$, per cui sia stato effettuato un pagamento diretto (valore 1)

5.2.2. Definizione dell'insieme di intervento
A seguito di questa operazione, l'insieme di mesi per i quali *è ragionevole ritenere l'esistenza* di dati mancanti (cioè la "non esistenza" del dato) è definita dalla soddisfazione dell'ulteriore vincolo:

→ dato il soddisfacimento della condizione precedente, gli importi *i* associati ai mesi *n* e *v* devono essere di un ordine di grandezza identico, cioè:

$$\left|(i_v - i_n)/i_n\right| \leq 0$$

Come si nota, in tal caso non è possibile utilizzare l'informazione proveniente dai trattamenti arretrati (non previsti per la tipologia specifica). L'aumento della

probabilità di errore associato alla riduzione degli elementi di validazione, ci induce a rendere più restrittivo il vincolo di inclusione.

5.2.3. Procedura di imputazione dei dati mancanti

Definito l'insieme di mesi per i quali si è identificata l'esistenza di dati mancanti, date le ipotesi assunte, l'imputazione degli importi avviene nel modo seguente:

→ ad ogni valore nullo si attribuisce un valore pari a quello linearmente interpolante i valori esterni alla/e non esistenza/e:

$$\hat{i}_m = i_v = i_n = i_v + \left[\left(i_n - i_v \right) \frac{m - v}{n - v} \right];$$

Le tavole che seguono forniscono informazioni sulla sensibilità della spesa al processo di correzione. Le tavole riportate seguono un livello di dettaglio crescente.

Tavola 5.4: Sensibilità della spesa al processo di correzione, per prestazione

	controllo no correzione		controllo e correzione		variazione % indotta	
prestazione*	spesa	n. mesi	spesa	n. mesi	spesa	n. mesi
010	2.362.820.366	1.986.146	2.362.820.366	1.986.146	0,000	0,000
040	29.649.885	29.692	29.649.885	29.692	0,000	0,000
141	2.410.809.515	1.487.344	2.411.433.956	1.489.597	0,026	0,151
160	4.684.434.891	5.529.327	4.766.976.529	5.679.670	1,762	2,719
170	805.850	850	805.850	850	0,000	0,000
200	48.998.778	75.785	48.998.778	75.785	0,000	0,000
221	688.804.502	140.471	688.804.502	140.471	0,000	0,000
281	90.338.359	162.797	90.338.359	162.797	0,000	0,000
Totale	10.316.662.146	9.412.412	10.399.828.225	9.565.008	0,806	1,621

Tavola 5.5: Sensibilità della spesa al processo di correzione, per settore di attività

	controllo no correzione		controllo e correzione		variazione % indotta	
ATECO	spesa	peso rel.	spesa	peso rel.	spesa	peso rel.
A	50.462.550	0,489	50.890.862	0,489	0,849	0,042
B	54.348.143	0,527	55.500.684	0,534	2,121	1,304
C	62.097.672	0,602	62.606.355	0,602	0,819	0,013
D	4.293.017.020	41,612	4.317.715.638	41,517	0,575	-0,229
E	52.163.083	0,506	52.599.188	0,506	0,836	0,030
F	1.369.499.808	13,275	1.386.982.215	13,337	1,277	0,467
G	1.307.123.628	12,670	1.319.411.743	12,687	0,940	0,133
H	464.385.204	4,501	468.975.869	4,509	0,989	0,181
I	780.962.814	7,570	787.123.557	7,569	0,789	-0,017
J	221.950.831	2,151	222.243.102	2,137	0,132	-0,669
K	862.715.832	8,362	870.474.374	8,370	0,899	0,092
L	14.717.952	0,143	14.877.592	0,143	1,085	0,276
M	75.611.211	0,733	76.225.345	0,733	0,812	0,006
N	360.339.357	3,493	361.446.872	3,476	0,307	-0,495
O	301.688.581	2,924	304.281.927	2,926	0,860	0,053
P	32.399	0,000	32.399	0,000	0,000	-0,800
Q	45.357.671	0,440	48.250.758	0,464	6,378	5,528
n.r.	188.390	0,002	189.745	0,002	0,719	-0,086
Totale	10.316.662.146	100,000	10.399.828.225	100,000	0,806	0,000

Tavola 5.6: Sensibilità della spesa al processo di correzione, per trattamento

			controllo no correzione		controllo e correzione		variazione % indotta	
prestazione*	trattamento*	pagamento*	spesa	n. mesi	spesa	n. mesi	spesa	n. mesi
010	01	0052	2.354.052.640	1.979.276	2.354.052.640	1.979.276	0,000	0,000
010	01	E778	8.767.726	6.870	8.767.726	6.870	0,000	0,000
040	02	0054	29.649.885	29.692	29.649.885	29.692	0,000	0,000
141	03	0053	1.836.597.471	695.954	1.836.597.471	695.954	0,000	0,000
141	03	E779	5.598.796	3.209	5.598.796	3.209	0,000	0,000
142	04	L050	342.355.164	423.972	342.355.164	423.972	0,000	0,000
143	04	L055	560.715	859	560.715	859	0,000	0,000
141	05	D800	182.779.409	300.225	182.779.409	300.225	0,000	0,000
141	05	D900	1.628.021	1.585	1.628.021	1.585	0,000	0,000
141	06	L053	1.261.369	1.425	1.273.465	1.448	0,959	1,614
141	06	L054	5.884.694	9.929	5.970.100	10.268	1,451	3,414
141	06	L056	21.505.096	34.089	21.880.305	35.486	1,745	4,098
141	06	L057	6.196.183	7.469	6.236.381	7.555	0,649	1,151
141	06	L058	6.442.597	8.628	6.554.129	9.036	1,731	4,729
160	07	0035	4.250.467.219	5.063.832	4.374.301.328	5.260.939	2,913	3,892
160	07	F240	712.033	219	712.033	219	0,000	0,000
160	07	H301	108.653	194	108.653	194	0,000	0,000
160	07	L036	433.075.755	465.073	391.783.284	418.309	-9,535	-10,055
160	07	T133	2.187	4	2.187	4	0,000	0,000
160	07	T151	10.532	4	10.532	4	0,000	0,000
160	07	T152	58.512	1	58.512	1	0,000	0,000
170	08	H017	805.850	850	805.850	850	0,000	0,000
200	09	L051	48.934.096	75.677	48.934.096	75.677	0,000	0,000
200	09	L052	64.682	108	64.682	108	0,000	0,000
221	10	0039	227.952.849	24.537	227.952.849	24.537	0,000	0,000
221	10	E200	2.699.330	836	2.699.330	836	0,000	0,000
221	10	E800	2.975.688	1.055	2.975.688	1.055	0,000	0,000
221	10	G400	205.649.796	112.863	205.649.796	112.863	0,000	0,000
221	10	V880	4.825	5	4.825	5	0,000	0,000
221	11	0040	230.674.781	681	230.674.781	681	0,000	0,000
221	11	F500	699.205	65	699.205	65	0,000	0,000
221	11	G600	3.321.806	97	3.321.806	97	0,000	0,000
221	11	G602	3.076.185	27	3.076.185	27	0,000	0,000
221	11	G603	11.621.915	299	11.621.915	299	0,000	0,000
221	11	G604	123.448	2	123.448	2	0,000	0,000
221	11	V890	4.674	4	4.674	4	0,000	0,000
281	12	R770	56.572.686	114.727	56.572.686	114.727	0,000	0,000
281	12	R771	21.316.019	37.131	21.316.019	37.131	0,000	0,000
281	12	R772	9.764.388	8.959	9.764.388	8.959	0,000	0,000
281	12	R773	1.135.068	715	1.135.068	715	0,000	0,000
281	12	R780	808.149	1.001	808.149	1.001	0,000	0,000
281	12	R781	173.452	109	173.452	109	0,000	0,000
281	12	R782	422.973	109	422.973	109	0,000	0,000
281	12	R783	145.624	46	145.624	46	0,000	0,000
Totale			10.316.662.146	9.412.412	10.399.828.225	9.565.008	0,806	1,621

Tavola 5.7: Sensibilità della spesa al processo di correzione, per prestazione e settore di attività

		controllo no correzione		controllo e correzione		variazione % indotta	
prestazione*	ATECO	spesa	peso rel.	spesa	peso rel.	spesa	peso rel.
141	A	9.920.406	0,411	9.928.682	0,412	0,083	0,058
141	B	147.725	0,006	147.725	0,006	0,000	-0,026
141	C	2.795.474	0,116	2.796.579	0,116	0,040	0,014
141	D	934.536.579	38,764	934.837.177	38,767	0,032	0,006
141	E	9.797.027	0,406	9.812.609	0,407	0,159	0,133
141	F	37.881.504	1,571	37.909.506	1,572	0,074	0,048
141	G	429.113.540	17,800	429.178.083	17,798	0,015	-0,011
141	H	96.433.299	4,000	96.451.008	4,000	0,018	-0,008
141	I	101.004.763	4,190	101.028.573	4,190	0,024	-0,002
141	J	168.977.040	7,009	169.005.551	7,009	0,017	-0,009
141	K	296.816.578	12,312	296.877.032	12,311	0,020	-0,006
141	L	4.605.146	0,191	4.608.648	0,191	0,076	0,050
141	M	39.583.287	1,642	39.596.771	1,642	0,034	0,008
141	N	181.395.857	7,524	181.432.454	7,524	0,020	-0,006
141	O	94.848.447	3,934	94.869.478	3,934	0,022	-0,004
141	P	1.670	0,000	1.670	0,000	0,000	-0,026
141	Q	2.868.643	0,119	2.869.880	0,119	0,043	0,017
141	n.r	82.530	0,003	82.530	0,003	0,000	-0,026
Totale		2.410.809.515	100,000	2.411.433.956	100,000	0,026	0,000

		controllo no correzione		controllo e correzione		variazione % indotta	
prestazione*	ATECO	spesa	peso rel.	spesa	peso rel.	spesa	peso rel.
160	A	29.574.375	0,631	29.994.411	0,629	1,420	-0,336
160	B	54.079.365	1,154	55.231.906	1,159	2,131	0,363
160	C	33.701.932	0,719	34.209.510	0,718	1,506	-0,252
160	D	1.822.882.075	38,914	1.847.280.095	38,752	1,338	-0,416
160	E	36.538.755	0,780	36.959.278	0,775	1,151	-0,601
160	F	829.624.143	17,710	847.078.548	17,770	2,104	0,336
160	G	556.467.310	11,879	568.690.882	11,930	2,197	0,427
160	H	228.421.333	4,876	232.994.289	4,888	2,002	0,236
160	I	401.937.644	8,580	408.074.577	8,560	1,527	-0,231
160	J	42.192.673	0,901	42.456.433	0,891	0,625	-1,117
160	K	365.709.204	7,807	373.407.292	7,833	2,105	0,337
160	L	6.402.898	0,137	6.559.036	0,138	2,439	0,665
160	M	18.566.912	0,396	19.167.562	0,402	3,235	1,448
160	N	85.047.437	1,816	86.118.355	1,807	1,259	-0,494
160	O	132.383.529	2,826	134.955.844	2,831	1,943	0,178
160	P	16.968	0,000	16.968	0,000	0,000	-1,732
160	Q	40.869.457	0,872	43.761.307	0,918	7,076	5,222
160	n.r	18.881	0,000	20.236	0,000	7,177	5,321
Totale		4.684.434.891	100,000	4.766.976.529	100,000	1,762	0,000

Come si può verificare, gli effetti del processo di correzione sulla spesa sono, in termini assoluti, alquanto contenuti. Risulta invece maggiore la sensibilità della distribuzione della spesa per forma di pagamento (tavola 5.6), per le ragioni descritte sopra.

6. Disaggregazione del dato

Sul dato così ottenuto viene effettuata una disaggregazione per codice ATECO ad 1 digit e per 6 classi dimensionali, definite in termini occupazionali. L'informazione necessaria a questa disaggregazione è, a meno della regionalizzazione, tutta disponibile all'interno della base di dati utilizzata. Essa non presenta criticità rilevanti per quanto riguarda la ricostruzione delle sezioni ATECO, effettuata in fase di condizionamento del dato alle definizioni ASIA (cnfr. par. 4.3).

È invece risultata problematica la definizione delle classi dimensionali, poiché il dato sugli occupati dipendenti a livello di impresa, proveniente da ASIA per le imprese ricostruite e dal servizio OCC per le rimanenti matricole (cfr. par. 4.4), risultava spesso logicamente incompatibile con il dato di spesa desumibile dalla fonte DM10 INPS[19]. In molti casi esso risultava incompatibile anche con altre informazioni provenienti dalla stessa fonte ASIA (fatturato).

[19] Per una descrizione di procedure alternative utilizzate si veda Cimino, Succi e Tuzi, (2000).

La tabella 6.1 fornisce una idea del problema di confrontabilità tra voci di spesa, fatturato e dato di occupazione.
Per ogni settore di attività economica e classe dimensionale calcolata sulla base dell'informazione proveniente da ASIA per le imprese ricostruite, si è calcolato il rapporto tra spesa aggregata per TMNP e fatturato, anch'esso proveniente da fonte ASIA. L'idea originaria era quella di avere una valutazione della variabilità dell'incidenza di spesa sul fatturato nelle classi definite. I risultati sono invece sorprendenti, poiché in alcune classi si registrano incidenze abbondantemente eccedenti il limite ragionevole dell'unità. In altri termini, in alcuni casi si ottiene che la spesa TMNP è superiore al fatturato delle imprese. Sebbene ciò sia ammissibile in una ottica di breve periodo e a livello di singola impresa, risulta di difficile giustificazione per incidenze medie calcolate su classi con numerosità spesso elevate.
I valori (medi!) di incidenza sospetti individuabili nella tavola 6.1, posto che il processo di ricostruzione delle imprese sia avvenuto correttamente, possono in effetti essere indotti sia da errori nel dato DM10, sia da errori nel dato di fatturato. La verifica effettuata confrontando il dato di fatturato con quello occupazionale, anziché risolvere gli inconvenienti, ha spesso sollevato ulteriori inconsistenze tra archivi.
L'analisi delle distribuzioni della spesa per dipendente ha prodotto problemi di confrontabilità analoghi. Ispezioni "manuali" sul dato elementare hanno indotto, nella maggioranza dei casi, all'identificazione di errori nel dato occupazionale piuttosto che in quello di spesa.
Ciò ha suggerito la costruzione di un modulo per la correzione dell'informazione sul numero di lavoratori dipendenti a livello di impresa, operante sulla base del riferimento di spesa per TMNP, che si assume consolidato a seguito delle procedure di controllo e correzione. A tale modulo viene dedicata una sezione specifica (par. 7).
Per quanto riguarda il dettaglio analitico a livello regionale si rileva che, sebbene sia possibile identificare la collocazione aziendale sul territorio, essa ha rilevanza esclusivamente contributiva, non fornendo alcuna indicazione sulla regione in cui vengono effettivamente riscossi i trattamenti. Da ciò deriva che i trattamenti erogati da aziende plurilocalizzate risultano attribuiti interamente alla regione in cui viene effettuata la denuncia di contribuzione e non consentono di rilevare informazioni sulla regione di residenza dei singoli beneficiari dei TMNP. La soluzione a tale problema richiede pertanto linee di ricerca mirate al momento ancora inesplorate[20].

[20] Per una breve rassegna sui possibili sviluppi futuri si veda il capitolo 10.

Tavola 6.1: Analisi delle inconsistenze tra fonti, per classe

ATECO	classe dim.*	media	numerosità	ATECO	classe dim.*	media	numerosità
A	1	1,83	1.099	I	4	0,06	1.643
A	2	0,27	224	I	5	0,06	891
A	3	0,11	114	I	6	0,17	509
A	4	0,32	35	J	1	1,65	5.117
A	5	1,58	10	J	2	0,41	1.049
A	6	0,02	1	J	3	7,29	740
B	1	0,26	993	J	4	1,04	340
B	2	0,27	309	J	5	1,06	346
B	3	0,40	156	J	6	0,19	405
B	4	1,09	42	K	1	0,11	54.841
B	5	2,55	16	K	2	0,03	13.473
B	6	0,10	5	K	3	0,14	8.041
C	1	0,03	1.178	K	4	0,13	2.687
C	2	0,02	715	K	5	0,54	1.602
C	3	0,02	576	K	6	1,15	853
C	4	0,03	128	L	1	0,05	190
C	5	0,00	34	L	2	0,10	79
C	6	0,03	7	L	3	0,28	72
D	1	0,31	88.981	L	4	0,97	16
D	2	0,19	50.073	L	5	0,57	11
D	3	0,05	46.953	L	6	20,10	7
D	4	0,03	14.754	M	1	0,07	1.711
D	5	0,04	7.569	M	2	2,06	685
D	6	0,04	3.477	M	3	0,56	593
E	1	0,01	330	M	4	0,79	229
E	2	0,22	210	M	5	0,62	67
E	3	0,01	244	M	6	0,05	14
E	4	0,09	140	N	1	0,09	13.235
E	5	0,01	92	N	2	0,08	2.090
E	6	0,05	84	N	3	0,17	1.658
F	1	0,12	85.893	N	4	1,31	820
F	2	0,93	23.392	N	5	0,30	654
F	3	0,06	12.137	N	6	0,12	289
F	4	0,10	2.315	O	1	0,30	18.341
F	5	150,10	737	O	2	0,08	3.848
F	6	0,02	199	O	3	0,08	2.104
G	1	0,02	112.659	O	4	0,23	684
G	2	0,01	30.461	O	5	0,16	331
G	3	0,02	16.024	O	6	1,20	165
G	4	0,01	3.638	P	1	0,37	7
G	5	0,01	1.258	P	2	0,04	1
G	6	0,02	575	P	3	0,04	1
H	1	0,16	36.291	P	4	0,00	0
H	2	0,14	8.440	P	5	26,72	1
H	3	0,03	4.202	P	6	0,00	0
H	4	0,23	866	Q	1	0,10	23
H	5	0,22	354	Q	2	0,14	10
H	6	0,08	152	Q	3	1,44	12
I	1	0,02	19.144	Q	4	1,06	4
I	2	0,03	7.121	Q	5	1,47	2
I	3	0,03	5.114	Q	6	2,41	1

Nota: la colonna ‘media’ rappresenta il valore medio calcolato nell’insieme [ATECO,Classe dimensionale] dei rapporti spesa TMNP e fatturato; analogamente la colonna ‘numerosità’ rappresenta il numero di imprese appartenenti all’insieme.

7. Procedure di correzione del dato occupazionale a livello di impresa

Abbiamo visto che il dato relativo ai dipendenti a livello di impresa desumibile dagli archivi ASIA e dalle elaborazioni del servizio OCC non risulta sempre compatibile con la dimensione di spesa per TMNP ottenuta dal consolidamento del dato DM10 INPS effettuato con la procedura sopra descritta.

Il criterio di compatibilità che è alla base di questa affermazione è stabilito da una ipotesi semplice di proporzionalità tra livello di spesa e dimensione di impresa. Sebbene l'utilizzo del riferimento di spesa sia questionabile (data la forte non sistematicità delle tipologie di spesa considerate), si è riscontrata una forte corrispondenza tra le incompatibilità ottenute adottando il riferimento alla spesa e quelle ottenute prendendo a riferimento variabili di impresa interne agli archivi di provenienza stessi (ASIA e OCC). In particolare, le incompatibilità tra spesa e dimensione spesso permangono se questa ultima è misurata in termini di fatturato piuttosto che in termini di occupazione.

Ciò induce a ritenere che la problematicità, almeno per i casi in cui manchi la proporzionalità occupazionale, sia rispetto alla spesa TMNP, sia rispetto al fatturato di impresa, debba essere attribuita al numero medio annuo di occupazione dipendenti a livello di impresa;

La definizione della disaggregazione per classi dimensionali richiede pertanto l'implementazione di una procedura di controllo e correzione dell'informazione sul numero di dipendenti.

Si assume che il dato di spesa aggregato per TMNP a livello di singola impresa sia vero ed esaustivo e che esista una relazione probabilistica diretta tra dimensione di spesa e dimensione di impresa, espressa in termini di numero di dipendenti. Ciò permette l'adozione a riferimento del valore di spesa per dipendente.

Esiste pertanto una distribuzione di importi erogati identificabile nei suoi momenti caratteristici.

Adottando una prospettiva tradizionale, la procedura di correzione più immediata è quella del calcolo dei primi due momenti delle distribuzioni degli importi per addetto e quindi della definizione di intervalli di confidenza per l'identificazione dei dati che non soddisfano il vincolo di inclusione. La correzione dei dati anomali avviene poi attraverso imputazione del valor medio per addetto e utilizzando quindi la spesa consolidata per la stima dei dipendenti.

Il meccanismo di correzione realizzato utilizza a grandi linee questo criterio di imputazione ma non il criterio di identificazione tradizionale accennato.

Le considerazioni alla base dell'abbandono di questa linea di approccio sono molteplici.

Dal punto di vista teorico, la definizione di un intervallo di confidenza in termini di percentuale di imprese attorno al valor medio presuppone un elemento soggettivo piuttosto forte, stante sostanzialmente nella definizione arbitraria di una misura di dispersione massima accettabile. Inoltre, sebbene tale definizione sia fondata nel caso di distribuzioni teoriche note, essa manca delle proprietà necessarie nel caso di distribuzioni effettive aventi caratteri strettamente peculiari.

È stato riscontrato in applicazioni sperimentali che l'utilizzo di questo approccio, data la forte asimmetria positiva, produce sensibili variazioni negative dei valori medi e mediani all'interno delle classi considerate e quindi

una forte alterazione in aumento del numero dei dipendenti nelle classi e nell'aggregato.
L'approccio qui proposto per l'identificazione dei dati anomali si basa esclusivamente sull'assunzione per cui il valor medio è tanto più rappresentativo dei valori della distribuzione quanto maggiore è la numerosità degli elementi della stessa e quanto minore è la loro dispersione (teorema del limite centrale).
Assumendo che esistano errori e che essi siano con maggiore probabilità negli estremi delle distribuzioni effettive, una procedura di correzione teoricamente ottimale è pertanto quella procedura che, dati i vincoli, produce la massima riduzione della dispersione con la minima perdita di numerosità.
Il meccanismo implementato è ispirato, sulla base dell'assunto fondamentale, alla procedura teorica ottimale ora definita, dato un criterio soggettivo per la definizione dell'ottimo (in sostanza un criterio di peso per numerosità e dispersione).
Per la definizione del criterio di ottimo si è optato per una formulazione il più possibile neutrale, quindi coerente con i dati. Gli elementi considerati sono pertanto le variazioni percentuali indotte dal processo di correzione ai momenti della distribuzione e alla numerosità dell'insieme di intervento, trattati in una ottica iterativa e simultanea.
Questo costituisce un elemento di distinzione rispetto all'approccio classico, che invece assume una logica statica di fissazione a priori di un criterio di selezione (numero di unità estreme sul totale) e correzione[21].
Alla descrizione tecnica vengono dedicate di seguito sezioni specifiche. È utile ribadire che concentreremo la nostra attenzione sulla definizione di valor medio degli importi per dipendente in termini di aspettativa matematica degli stessi, con associate misure di dispersione e di simmetria. Si assumerà inoltre che le caratteristiche salienti degli importi per addetto possano essere catturate nelle due dimensioni controllate nell'analisi: settore di attività economica e classe dimensionale espressa in termini di numero di dipendenti.

7.1. Identificazione iterativa dei dati anomali

Dato il settore di attività economica *m*-esimo ($m = 1,\ldots,17$) e la classe dimensionale *n*-esima ($n = 1,\ldots,6$), vengono in una prima fase calcolati numerosità, momenti, coefficiente di variazione e valore mediano per ognuna delle 102 distribuzioni specifiche. Sia $D_{i=0}$ l'insieme delle distribuzioni originarie, cioè l'insieme delle (17×6) distribuzioni all'iterazione $i = 0$.
Questo insieme è cruciale per l'esito del processo di identificazione dei dati anomali poiché il criterio di ottimalità farà riferimento ad esso nella definizione dei pesi variabili utilizzati nelle iterazioni del processo di correzione.
Per ogni iterazione *i* del processo e per ogni distribuzione identificata dalla coppia (m,n), l'esistenza di dati anomali di impresa O_i è definita dalla mancata soddisfazione della seguente disuguaglianza:

[21] Per la descrizione di un metodo di identificazione degli outlier basato su scelte soggettive dei valori di soglia si veda Hidiroglou e Berthelot, 1986. Per una metodologia di identificazione alternativa a quella a soglie fisse si veda Gismondi, 2002. Per una rassegna dei metodi consolidati si veda Barnett e Lewis, 1994.

$$O_i = \frac{\left| x_{m,n} - E(x_{m,n})_{i=0} + \left[E(x_{m,n})_{i=0} - Me(x_{m,n})_{i=0} \right] \right|}{(\sigma_{m,n})_{i=0}} \leq (K_{m,n})_i \quad \forall m,n,i$$

dove $x_{m,n}$ è l'elemento della distribuzione (spesa per dipendente) relativa alla sezione ATECO m e alla classe dimensionale n, $E(x_{m,n})_{i=0}$ è la sua media calcolata alla prima iterazione $i=0$, $Me(x_{m,n})_{i=0}$, è il valore mediano della distribuzione calcolato alla prima iterazione $i=0$, $(\sigma_{m,n})_{i=0}$ è lo scarto quadratico della distribuzione m, j per la stessa iterazione $i=0$.

Il criterio di discriminazione espresso dal valore $(K_{m,n})_i$ varia per ogni iterazione del processo. Dal momento che la convergenza al criterio di ottimo potrebbe avvenire con velocità diverse nelle diverse distribuzioni m, j, esso viene espresso anche rispetto alle due dimensioni considerate, tenendo tuttavia presente che, per l'iterazione *i*-esima e nel caso il processo sia attivo per tutte le distribuzioni attivate, il parametro di discriminazione assume valori comuni.

Il significato dell'espressione scritta sopra, se considerata sotto una prospettiva statica, cioè prescindendo dall'iterazione *i* della stessa, è in sostanza quello di una generica definizione di un intervallo di confidenza per le distribuzioni m, j, dove si è voluto controllare per l'eventuale asimmetria.

La modificazione del processo di selezione rispetto alle componenti di asimmetria è espressa dal termine in parentesi quadra. Il baricentro della selezione diventa pertanto la mediana.

Il dispositivo di selezione implementato, a differenza delle soluzioni statiche, non richiede una definizione arbitraria del parametro di discriminazione *K*. Alla iterazione 0 si definisce un valore grande a piacere per $K_{i=0}$ ed esso, ad ogni iterazione, viene ridotto di un ordine di grandezza non nullo piccolo a piacere (*passo* dell'ottimizzazione), fino a terminazione del processo

$$K_i = K_{i=0} - \sum_{i=1}^{n} step_i$$

dove $step_i$ è la dimensione del passo di ottimizzazione (analogo del passo di integrazione negli algoritmi di ottimizzazione non lineare).

È evidente che, dato un valore costante di *passo*, maggiore è il valore di *start* per $K_{i=0}$, maggiore è il numero di iterazioni *n*:

$$n_{m,n} = \frac{K_{i=0} - (K_i^*)_{m,n}}{step_i}$$

dove $(K_i^*)_{m,n}$ è il valore assunto dal parametro di discriminazione alla terminazione del processo iterativo, cioè ad ottimizzazione avvenuta.

Come si nota, il criterio di discriminazione adottato non è specificato a priori, ma è il risultato dei criteri di terminazione adottati, che nella fattispecie sono anche criteri di ottimo. La sezione di seguito discute la questione dei criteri di terminazione.

7.2. Criteri di terminazione (o di stop) e criteri di ottimalità

Se $K_{i=0}$ è molto grande vengono prodotte molte iterazioni vuote (ad effetto nullo), a scapito dell'efficienza. Allo stesso modo, riducendo il valore di passo, si aumentano il numero di iterazioni potenziali.

La questione della definizione del passo ottimale è una questione cruciale molto dibattuta e che sostanzialmente non trova soluzioni univoche. Per l'analisi numerica su superfici "levigate" la riduzione del passo garantisce maggiore precisione di calcolo (nel caso limite si opta per un valore pari a quello della precisione numerica delle macchine), ma ciò non appare essere sempre la scelta migliore. Nel caso di superfici (nel nostro caso distribuzioni) ad andamento molto variabile, l'adozione di un passo molto contenuto può indurre a soluzioni solo localmente ottimali, cioè alla convergenza precoce del processo.

Nelle applicazioni effettuate abbiamo adottato un passo costante di dimensione 5×10^{-4}, ed un parametro di $start = 3$ che, alla prima iterazione, generava selezioni vuote. La scelta ha prodotto risultati soddisfacenti in termini di efficienza e di risultato. È evidente che tale qualificazione fa riferimento alle ipotesi e alle misure prese in considerazione. La questione della definizione di un passo variabile endogeno al processo di discriminazione verrà affrontata a risultati consolidati.

Nella logica iterativa qui assunta i criteri di ottimo si risolvono in criteri di terminazione, o di stop. Si tratta di definire delle condizioni la cui soddisfazione definisce allo stesso tempo la terminazione del processo e la sua ottimalità.

Ad ogni iterazione del processo gli elementi descrittivi delle m,n distribuzioni D_i vengono nuovamente calcolati. Dato l'assunto fondamentale si è stabilito che una procedura di correzione teoricamente ottimale è quella procedura che, dati i vincoli, produca la massima riduzione della dispersione con la minima perdita di numerosità.

Sulla base di ciò viene stabilito il seguente criterio di terminazione:

→ *la discriminazione è terminata se, all'iterazione i-esima, la variazione percentuale (pesata) indotta alla dispersione della distribuzione rispetto all'iterazione precedente è inferiore o uguale alla variazione percentuale (pesata) in modulo indotta alla numerosità degli elementi della stessa distribuzione*;

→ *ad ogni iterazione i termini della disuguaglianza vengono pesati per il rispettivo rapporto rispetto ai loro valori originari o per l'inversa di tale rapporto;*

→ *la soluzione ottimale è quella immediatamente precedente l'iterazione in cui le condizioni stabilite ai punti precedenti vengono soddisfatte, quindi quella associata ad un* $K^*_{i=(n-1)}$.

In formule:

$$n_{m,n} = n^*_{m,n};\ \ (K_{m,n})_i = (K^*_{m,n})_{i=(n-1)}$$

$$\text{se} \begin{cases} (\alpha_{m,n})_i < (\beta_{m,n})_i \\ (\alpha_{m,n})_i = \dfrac{(\sigma_{m,n})_i - (\sigma_{m,n})_{i-1}}{(\sigma_{m,n})_{i-1}} \cdot \dfrac{(\sigma_{m,n})_i}{(\sigma_{m,n})_{i=0}} \\ (\beta_{m,n})_i = \dfrac{\left|(N_{m,n})_i - (N_{m,n})_{i-1}\right|}{(N_{m,n})_{i-1}} \cdot \dfrac{(N_{m,n})_{i=0}}{(N_{m,n})_i} \end{cases}$$

dove $(N_{m,n})_i$ è la numerosità della distribuzione m,n-esima alla *i*-esima iterazione.

Il processo descritto si riduce sostanzialmente alla ricerca dell'uguaglianza nelle differenze prime di numerosità e dispersione calcolate rispetto a variazioni del parametro di sensibilità per la discriminazione.

I termini di correzione (pesi) che appaiono nella definizione della disuguaglianza non sono strettamente necessari alla convergenza del processo e il loro uso è facoltativo. Essi costituiscono l'elemento soggettivo del criterio di ottimo, che quindi assume i connotati di un criterio informativo di selezione. Si assume che quanto maggiore è l'allontanamento dalle distribuzioni originarie $D_{i=0}$ indotto dal processo di correzione, tanto maggiore è il rischio di incorrere in errori di selezione, o di esclusione.

L'introduzione del rapporto tra valori calcolati all'iterazione *i*-esima e valori originari per la deviazione standard e dell'inverso di tale rapporto per la numerosità introduce un elemento di non linearità che rende sempre maggiori i guadagni necessari per il proseguimento del processo al crescere della distanza dai valori delle distribuzioni iniziali.

Come si nota il criterio di terminazione adottato corrisponde al criterio di ottimo assunto, a meno dell'elemento – opzionale – di peso soggettivo.

Si deriva intuitivamente che la procedura implementata termina la selezione quando incontra il "*corpo*" delle distribuzioni effettive. Se la distribuzione è fortemente asimmetrica l'algoritmo tende a selezionare di più sul lato più lungo della distribuzione. In caso di distribuzioni simmetriche anche la selezione è simmetrica.

Abbiamo testato la procedura su dati estratti da distribuzioni normali standardizzate generate artificialmente ottenendo in media una selezione simmetrica per il 13,5% delle unità. Escludendo i termini di correzione, la procedura termina le iterazioni in corrispondenza del punto di flesso della distribuzione normale standardizzata, cioè discrimina gli elementi esterni all'intervallo $[-1,1]$ di deviazione standard.

In caso di distribuzioni molto piatte può tuttavia accadere che l'algoritmo non incontri le condizioni di terminazione fino a svuotamento totale delle classi (la variazione indotta alla dispersione è sempre superiore alla riduzione indotta alla numerosità). In questi casi, più probabili se si prescinde dall'utilizzo del peso soggettivo, l'algoritmo si ferma quando incontra la seguente ulteriore condizione di terminazione:

$$\frac{(N_{m,n})_i}{(N_{m,n})_{i=0}} \leq R_{\min}$$

L’adozione del criterio scritto sopra è del tutto arbitraria e sostanzialmente la sua funzione è quella di segnalare la mancata convergenza del processo. Infatti, distribuzioni del tipo configurato presentano valori di dispersione e di Kurtosi tali da scoraggiare l’uso della media come valore atteso e quindi come valore di correzione.
Una condizione di terminazione ulteriore è la seguente:

$$(N_{m,n})_{i=0} \leq N_{\min}$$

Anche in tal caso, la definizione del criterio è arbitraria e convenzionale e il suo significato è che si escludono dai processi di correzione quelle distribuzioni la cui bassa numerosità renderebbe irragionevoli le ipotesi operative alla base della procedura.
La tavola 7.1 contiene i risultati del processo automatico di identificazione dei dati anomali. Vengono riportati i valori iniziali e di terminazione per media, mediana, asimmetria, deviazione standard, numerosità e coefficiente di variazione, ripetuti per le 17×6 distribuzioni alle quali si applica il processo di identificazione. Come si nota, a fronte di forti guadagni in dispersione e asimmetria, le cadute in numerosità appaiono relativamente moderate ed in media pari a circa il 7% delle numerosità originarie. I valori mediani subiscono alterazioni contenute.

Tab. 7.1: Effetti sulle distribuzioni originarie dell'applicazione algoritmo di identificazione dei dati anomali

classe	Media	Mediana	Asimm.	dev. Std	N	CV	classe	Media	Mediana	Asimm.	dev. Std	N	CV
[A , 1] :	1832.0	1184.0	2.55	2003.5	2809	1.09	[E , 2] :	1006.4	564.7	6.41	1581.8	143	1.57
	1431.7	1063.9	0.90	1224.3	2606	0.86		807.7	520.7	1.14	759.0	135	0.94
[A , 2] :	898.9	595.3	2.39	961.7	328	1.07	[E , 3] :	754.5	546.0	1.33	658.1	204	0.87
	713.1	521.1	0.87	603.2	305	0.85		637.5	485.4	0.85	488.0	189	0.77
[A , 3] :	787.8	548.4	2.11	752.8	221	0.96	[E , 4] :	812.7	650.8	1.43	606.2	290	0.75
	669.7	516.3	0.82	524.4	208	0.78		719.2	604.1	0.75	455.7	273	0.63
[A , 4] :	726.2	537.7	2.04	635.4	122	0.87	[E , 5] :	737.0	694.7	1.38	444.5	68	0.60
	620.7	512.7	0.76	423.9	114	0.68		679.2	691.8	0.16	316.4	64	0.47
[A , 5] :	419.4	383.6	0.52	389.2	8	0.93	[E , 6] :	572.2	476.2	1.30	413.0	60	0.72
	419.4	383.6	0.52	389.2	8	0.93		490.7	452.1	0.45	283.2	54	0.58
[A , 6] :	1570.9	1570.9	0.00	976.0	2	0.62	[F , 1] :	1895.9	1287.0	2.58	2023.6	103858	1.07
	1570.9	1570.9	0.00	976.0	2	0.62		1481.6	1132.0	0.85	1276.0	95983	0.86
[B , 1] :	2547.7	2051.7	4.31	2444.0	2661	0.96	[F , 2] :	1823.6	1510.9	1.34	1499.2	18697	0.82
	2167.6	1940.0	0.58	1469.3	2507	0.68		1548.4	1371.1	0.52	1107.5	17290	0.72
[B , 2] :	2233.4	2030.3	0.70	1353.4	423	0.61	[F , 3] :	1828.1	1565.0	1.55	1390.9	12009	0.76
	2132.4	2003.6	0.33	1206.4	410	0.57		1603.7	1469.6	0.46	1049.1	11219	0.65
[B , 3] :	2527.8	2378.4	0.62	1318.8	243	0.52	[F , 4] :	1948.8	1705.3	1.54	1292.8	4909	0.66
	2437.7	2347.5	0.18	1084.2	227	0.44		1781.9	1637.9	0.47	1019.5	4655	0.57
[B , 4] :	2451.3	2381.1	0.98	1403.0	144	0.57	[F , 5] :	2165.8	1824.9	7.26	2144.4	218	0.99
	2319.2	2367.3	-0.13	998.1	129	0.43		1911.4	1753.5	0.50	1031.6	209	0.54
[B , 5] :	2080.0	2285.7	0.17	1435.6	15	0.69	[F , 6] :	2209.9	1735.7	4.15	2251.1	81	1.02
	1764.4	1630.8	0.13	1236.3	13	0.70		1841.9	1693.9	0.73	1073.6	76	0.58
[B , 6] :	892.8	54.1	0.75	1696.6	4	1.90	[G , 1] :	1639.5	842.8	2.99	2181.6	130940	1.33
	892.8	54.1	0.75	1696.6	4	1.90		1169.0	712.5	1.15	1209.6	120940	1.03
[C , 1] :	2138.3	1524.3	2.15	2089.6	1429	0.98	[G , 2] :	1027.8	680.1	1.61	1045.4	23062	1.02
	1748.3	1422.4	0.81	1374.1	1325	0.79		813.8	601.2	0.88	718.0	21204	0.88
[C , 2] :	2021.0	1694.3	1.65	1606.2	558	0.79	[G , 3] :	964.2	732.8	1.47	843.2	15333	0.87
	1736.8	1594.0	0.56	1131.1	520	0.65		802.0	660.2	0.71	602.0	14167	0.75
[C , 3] :	2002.2	1664.7	1.65	1489.6	568	0.74	[G , 4] :	928.1	780.5	1.57	654.8	7275	0.71
	1772.5	1569.9	0.53	1090.3	537	0.62		827.2	743.8	0.62	490.7	6874	0.59
[C , 4] :	1931.4	1359.8	1.56	1602.0	258	0.83	[G , 5] :	931.4	789.9	8.31	684.6	536	0.74
	1545.9	1232.5	1.22	1039.7	233	0.67		871.6	786.2	0.82	397.3	521	0.46
[C , 5] :	1500.6	1710.5	-0.23	841.4	6	0.56	[G , 6] :	995.6	979.7	0.86	438.1	244	0.44
	1500.6	1710.5	-0.23	841.4	6	0.56		944.3	944.0	0.10	338.7	224	0.36
[C , 6] :	992.2	637.5	0.60	1099.4	4	1.11	[H , 1] :	1480.2	812.7	3.04	1907.3	41479	1.29
	992.2	637.5	0.60	1099.4	4	1.11		1065.1	707.5	1.13	1037.7	38289	0.97
[D , 1] :	1484.1	758.3	166.82	6094.4	106454	4.11	[H , 2] :	991.2	670.6	1.62	983.8	7835	0.99
	1147.4	688.8	1.33	1214.2	101775	1.06		786.8	593.8	0.89	671.6	7229	0.85
[D , 2] :	1016.8	676.0	21.59	1261.5	40043	1.24	[H , 3] :	984.4	758.9	1.40	821.1	4896	0.83
	805.8	612.5	0.89	689.1	36850	0.86		840.0	705.0	0.66	605.3	4571	0.72
[D , 3] :	1041.1	790.5	86.97	1563.5	43101	1.50	[H , 4] :	1036.9	911.8	1.31	705.3	2253	0.68
	867.5	734.1	0.72	622.1	40169	0.72		930.5	854.8	0.48	542.0	2126	0.58
[D , 4] :	1088.3	890.9	20.99	1059.6	30250	0.97	[H , 5] :	959.2	945.2	0.22	467.9	146	0.49
	951.8	855.5	0.72	545.7	28710	0.57		939.1	938.3	0.07	390.8	132	0.42
[D , 5] :	1066.8	906.2	10.07	874.8	3542	0.82	[H , 6] :	1295.5	1239.9	0.84	470.3	78	0.36
	960.6	887.3	0.86	438.7	3415	0.46		1253.4	1238.6	0.09	288.9	67	0.23
[D , 6] :	1037.2	899.2	10.51	897.7	1425	0.87	[I , 1] :	1702.1	1021.3	3.76	2031.6	22762	1.19
	939.9	880.8	0.80	402.1	1382	0.43		1290.0	906.3	1.02	1194.7	21029	0.93
[E , 1] :	1341.0	772.0	3.41	1672.4	437	1.25	[I , 2] :	1263.0	938.3	1.71	1167.5	5803	0.92
	982.3	668.8	1.09	892.1	402	0.91		1036.3	844.0	0.75	820.2	5358	0.79

segue

Nota: la prima riga relativa alla classe $[m,n]$, (m = sezione ATECO, n = classe dimensionale), fa riferimento alle distribuzioni originarie, la seconda alle distribuzioni ottenute a terminazione del processo.

Tab. 7.1 segue

classe	Media	Mediana	Asimm.	dev. Std	N	CV	classe	Media	Mediana	Asimm.	dev. Std	N	CV
[I , 3] :	1268,1	992,4	1,98	1091,8	4742	0,86	[M , 4] :	762,0	648,7	1,20	548,0	847	0,72
	1065,1	**913,0**	**0,69**	**764,9**	**4408**	**0,72**		**675,1**	**624,2**	**0,46**	**419,8**	**797**	**0,62**
[I , 4] :	1154,6	929,1	1,58	881,0	3390	0,76	[M , 5] :	914,0	742,1	0,64	443,8	43	0,49
	990,0	**876,9**	**0,75**	**619,5**	**3162**	**0,63**		**912,1**	**778,3**	**0,67**	**363,3**	**38**	**0,40**
[I , 5] :	1082,7	895,1	2,12	788,1	480	0,73	[M , 6] :	872,1	866,5	-0,17	348,0	9	0,40
	949,5	**854,0**	**0,71**	**540,0**	**451**	**0,57**		**872,1**	**866,5**	**-0,17**	**348,0**	**9**	**0,40**
[I , 6] :	1273,9	915,3	1,00	949,4	254	0,75	[N , 1] :	1623,9	762,0	3,51	2285,2	16905	1,41
	1195,3	**893,4**	**0,73**	**834,3**	**245**	**0,70**		**1156,4**	**634,2**	**1,17**	**1260,4**	**15620**	**1,09**
[J , 1] :	1998,7	1055,5	3,37	2622,5	6243	1,31	[N , 2] :	1051,5	694,9	1,72	1104,3	2217	1,05
	1449,1	***915,0***	***1,13***	***1451,1***	***5779***	***1,00***		***822,1***	***604,8***	***0,85***	***737,0***	***2028***	***0,90***
[J , 2] :	1040,2	606,1	1,86	1150,1	831	1,11	[N , 3] :	1121,1	912,3	*1,56*	963,0	1869	0,86
	807,9	**531,9**	**1,09**	**763,1**	**764**	**0,94**		**943,5**	**834,5**	**0,54**	**680,7**	**1734**	**0,72**
[J , 3] :	911,8	540,0	2,49	998,0	672	1,09	[N , 4] :	1189,2	1101,3	1,20	703,6	2157	0,59
	740,8	**489,2**	**0,80**	**671,0**	**631**	**0,91**		**1112,6**	**1074,0**	**0,35**	**572,1**	**2059**	**0,51**
[J , 4] :	726,6	553,8	1,79	650,3	845	0,89	[N , 5] :	1411,6	1272,4	15,61	1492,6	373	1,06
	595,9	**505,3**	**0,64**	**439,5**	**777**	**0,74**		**1343,1**	**1272,4**	**0,55**	**530,8**	**371**	**0,40**
[J , 5] :	621,8	555,0	1,03	425,7	234	0,68	[N , 6] :	1448,5	1416,6	0,66	519,3	132	0,36
	570,2	**546,3**	**0,29**	**346,4**	**222**	**0,61**		**1374,0**	**1385,3**	**0,12**	**341,9**	**115**	**0,25**
[J , 6] :	474,5	430,7	0,94	329,2	244	0,69	[O , 1] :	1513,4	640,0	11,44	2416,3	24345	1,60
	419,9	**409,6**	**0,32**	**257,9**	**226**	**0,61**		**1043,8**	**534,0**	**1,28**	**1170,1**	**22452**	**1,12**
[K , 1] :	1901,1	965,0	7,65	2616,1	81715	1,38	[O , 2] :	930,1	573,9	1,80	1020,1	3534	1,10
	1420,7	**840,0**	**1,04**	**1447,0**	**75298**	**1,02**		**722,6**	**508,3**	**0,95**	**678,4**	**3224**	**0,94**
[K , 2] :	982,8	567,4	2,63	1125,2	10478	1,14	[O , 3] :	966,1	695,8	2,17	957,6	2495	0,99
	762,9	**491,0**	**0,98**	**749,3**	**9703**	**0,98**		**785,2**	**638,4**	**0,78**	**637,3**	**2287**	**0,81**
[K , 3] :	957,7	699,9	1,76	919,2	7818	0,96	[O , 4] :	931,6	742,3	2,20	842,9	1756	0,90
	781,6	**630,8**	**0,72**	**635,4**	**7203**	**0,81**		**765,6**	**670,6**	**0,66**	**549,6**	**1619**	**0,72**
[K , 4] :	1005,1	809,8	13,55	1048,7	5756	1,04	[O , 5] :	1156,8	962,0	1,06	932,9	190	0,81
	838,0	**761,8**	**0,61**	**558,6**	**5365**	**0,67**		**972,5**	**925,1**	**0,61**	**684,1**	**172**	**0,70**
[K , 5] :	1026,1	840,9	2,19	800,3	830	0,78	[O , 6] :	989,1	772,6	3,95	1192,4	88	1,21
	853,9	**803,4**	**0,75**	**470,8**	**768**	**0,55**		**810,9**	**713,6**	**0,50**	**651,3**	**81**	**0,80**
[K , 6] :	952,5	834,8	1,76	677,1	376	0,71	[Q , 1] :	2607,6	2011,0	1,06	2268,3	576	0,87
	830,9	**786,9**	**0,47**	**462,9**	**349**	**0,56**		**2249,1**	**1749,0**	**0,66**	**1794,3**	**540**	**0,80**
[L , 1] :	1398,3	371,2	10,73	6022,6	135	4,31	[Q , 2] :	759,1	515,0	1,07	813,1	33	1,07
	803,1	**385,3**	**1,70**	**934,2**	**123**	**1,16**		**695,8**	**550,8**	**0,66**	**642,7**	**28**	**0,92**
[L , 2] :	751,8	451,5	1,51	882,6	104	1,17	[Q , 3] :	633,9	607,3	0,45	512,0	19	0,81
	526,2	**370,5**	**1,06**	**519,9**	**92**	**0,99**		**527,5**	**549,0**	**0,33**	**420,2**	**17**	**0,80**
[L , 3] :	871,2	632,7	1,65	857,9	137	0,98	[Q , 4] :	589,1	360,3	2,83	894,3	16	1,52
	731,4	**602,3**	**0,60**	**594,3**	**127**	**0,81**		**400,1**	**360,3**	**0,76**	**262,3**	**14**	**0,66**
[L , 4] :	849,4	684,3	0,93	722,9	161	0,85	[+ , 1] :	1677,4	961,0	3,77	2152,5	374	1,28
	725,5	**546,2**	**0,56**	**559,7**	**147**	**0,77**		**1253,1**	**864,6**	**0,99**	**1176,6**	**338**	**0,94**
[L , 5] :	589,5	433,7	0,67	496,2	35	0,84	[+ , 2] :	1420,2	529,0	7,77	4545,6	77	3,20
	539,8	**388,2**	**0,45**	**421,2**	**32**	**0,78**		**898,5**	**538,3**	**1,55**	**965,1**	**72**	**1,07**
[L , 6] :	537,7	442,2	0,19	474,4	8	0,88	[+ , 3] :	1108,6	933,9	1,05	931,5	66	0,84
	537,7	**442,2**	**0,19**	**474,4**	**8**	**0,88**		**921,9**	**867,8**	**0,51**	**678,2**	**59**	**0,74**
[M , 1] :	1282,4	535,3	23,15	2557,7	4591	1,99	[+ , 4] :	918,7	791,3	1,30	701,7	50	0,76
	892,0	**466,0**	**1,24**	**1000,5**	**4209**	**1,12**		**807,0**	**776,6**	**0,34**	**486,0**	**46**	**0,60**
[M , 2] :	813,0	455,6	1,85	951,8	1696	1,17	[+ , 5] :	916,9	916,9	0,00	55,7	2	0,06
	614,4	**377,8**	**1,05**	**631,7**	**1546**	**1,03**		**916,9**	**916,9**	**0,00**	**55,7**	**2**	**0,06**
[M , 3] :	789,5	586,2	1,34	736,3	1233	0,93	[+ , 6] :	870,4	870,4	0,00	115,5	2	0,13
	637,8	**516,6**	**0,72**	**521,9**	**1124**	**0,82**		**870,4**	**870,4**	**0,00**	**115,5**	**2**	**0,13**

Nota: $[0,3]$ è l'intervallo di ricerca del parametro di discriminazione K; il passo di scansione è $0,005$; $N_{\min} = 10$; $R_{\min} = 0,5$;

7.3. Correzione dati anomali

I valori considerati per la donazione del dato per dipendente sono esclusivamente quelli calcolati al livello di sezione ATECO poiché l'informazione rispetto alla classe dimensionale non può essere considerata disponibile prima del completamento della procedura.

A prescindere dalla soluzione adottata, che descriviamo in seguito, il recupero del dato occupazionale viene effettuato utilizzando la seguente:

$$(\hat{D}ip)_{O\in m} = \frac{(S_{TMNP})_{O\in m}}{E(S_{TMNP}/Dip)_m}$$

dove $(\hat{D}ip)_{O\in m}$ è il numero di dipendenti da stimare per le imprese definite anomale, $(S_{TMNP})_{O\in m}$ è la spesa aggregata per TMNP nelle stesse imprese e $E(S_{TMNP}/Dip)_m$ è l'aspettativa matematica dell'importo di spesa per addetto nella sezione ATECO a cui appartiene l'impresa definita anomala $O \in m$.

La procedura è stata testata su diverse ipotesi di donazione, impropriamente, su diversi modi di calcolo dell'aspettativa matematica.

La soluzione più immediata è quella dell'utilizzo della media calcolata sulla numerosità originaria associata ad ogni sezione ATECO. Ciò significa il recupero, ai fini della donazione, delle imprese che il processo di identificazione ha definito come anomale.

Una seconda soluzione è quella dell'utilizzo della media calcolata sugli insiemi al netto delle imprese anomale.

Una via alternativa a quella basata sulle medie campionarie, indipendentemente dal campo di osservazione (insieme originario vs insieme derivato dalla selezione) è l'utilizzo del dato stimato attraverso analisi di regressione della relazione tra spesa TMNP e dipendenti all'interno delle diverse sezioni ATECO. Data la presenza di forti componenti eteroschedastiche, la stima viene condotta utilizzando uno stimatore GLS (minimi quadrati generalizzati).

La sperimentazione ha dimostrato che i risultati sono simili a quelli ottenibili attraverso l'utilizzo di medie robuste.

Nel caso dell'utilizzo dello stimatore GLS la relazione scritta sopra si traduce nella seguente:

$$(\hat{D}ip)_{O\in m} = \frac{(S_{TMNP})_{O\in m}}{(\hat{s}_m)_{GLS}}$$

dove $(\hat{s}_m)_{GLS}$ è il parametro della regressione tra spesa TMNP e dipendenti a livello di impresa stimata per ogni sezione ATECO *m*.

La tavola 7.2 contiene i risultati del processo di identificazione e correzione del dato occupazionale, effettuato adottando il secondo approccio descritto, centrato sulle medie delle distribuzioni "pulite". Vengono riportati i valori originari, quelli a terminazione del processo di correzione e gli scostamenti indotti per la numerosità delle distribuzioni (matricole), i dipendenti e la spesa ad esse associata. Come si nota, il processo di correzione produce un aumento dell'occupazione complessiva ma nessun aumento della spesa. Con riferimento a questa ultima, le alterazioni vengono prodotte solo all'interno del singolo settore di attività economica, nella distribuzione per classi di età, ma non nell'aggregato.

Tab. 7.2: Effetti del processo di correzione, per sezione e classe dimensionale

ATECO	cl. dim.	N i = 0	N i = i*	Variaz. N	Dip. i = 0	Dip. i = i*	Variaz. Dip	Spesa per i = 0	Spesa per i = i*	Variaz. spesa
A	1	2.809	2.731	-0,028	5.594	5.747	0,027	8.922.164	7.602.160	-0,148
A	2	328	347	0,058	2.321	2.475	0,066	2.046.535	2.006.151	-0,020
A	3	221	240	0,086	2.996	3.243	0,082	2.378.216	2.527.522	0,063
A	4	122	153	0,254	4.563	5.659	0,240	3.119.102	3.549.293	0,138
A	5	8	15	0,875	1.034	1.973	0,908	458.203	1.054.441	1,301
A	6	2	3	0,500	758	1.045	0,379	1.285.988	1.464.177	0,139
A	*tot 1-6*	*3.490*	*3.489*	*0,000*	*17.266*	*20.142*	*0,167*	*18.210.208*	*18.203.744*	*0,000*
B	1	2.661	2.583	-0,029	7.329	7.216	-0,015	17.991.370	15.993.704	-0,111
B	2	423	463	0,095	3.007	3.312	0,101	6.757.032	7.192.882	0,065
B	3	243	270	0,111	3.066	3.444	0,123	7.764.764	8.255.321	0,063
B	4	144	152	0,056	5.837	6.026	0,032	14.591.484	14.702.498	0,008
B	5	15	17	0,133	2.175	2.585	0,189	4.232.730	4.683.364	0,106
B	6	4	5	0,250	1.394	1.682	0,207	992.133	1.501.744	0,514
B	*tot 1-6*	*3.490*	*3.490*	*0,000*	*22.808*	*24.265*	*0,064*	*52.329.513*	*52.329.513*	*0,000*
C	1	1.429	1.377	-0,036	3.826	3.739	-0,023	7.954.622	6.601.896	-0,170
C	2	558	541	-0,030	4.054	3.929	-0,031	8.199.342	6.823.788	-0,168
C	3	568	568	0,000	7.571	7.614	0,006	15.112.816	13.443.810	-0,110
C	4	258	318	0,233	8.466	11.473	0,355	15.320.809	17.983.596	0,174
C	5	6	15	1,500	817	1.934	1,367	1.265.431	2.999.930	1,371
C	6	4	4	0,000	6.196	6.196	0,000	3.862.001	3.862.001	0,000
C	*tot 1-6*	*2.823*	*2.823*	*0,000*	*30.930*	*34.885*	*0,128*	*51.715.021*	*51.715.021*	*0,000*
D	1	106.454	103.629	-0,027	282.817	281.407	-0,005	365.196.906	306.741.555	-0,160
D	2	40.043	38.276	-0,044	292.592	280.078	-0,043	297.625.337	230.840.727	-0,224
D	3	43.101	41.669	-0,033	580.704	561.704	-0,033	605.668.315	496.265.428	-0,181
D	4	30.250	35.368	0,169	1.164.477	1.403.334	0,205	1.266.166.796	1.325.535.027	0,047
D	5	3.542	4.178	0,180	532.192	623.812	0,172	567.107.798	597.653.858	0,054
D	6	1.425	1.695	0,189	1.008.734	1.272.521	0,262	1.087.409.616	1.232.138.173	0,133
D	*tot 1-6*	*224.815*	*224.815*	*0,000*	*3.861.516*	*4.422.856*	*0,145*	*4.189.174.768*	*4.189.174.768*	*0,000*
E	1	437	418	-0,043	1.227	1.198	-0,024	1.508.882	1.180.256	-0,218
E	2	143	142	-0,007	1.055	1.048	-0,007	1.050.460	866.515	-0,175
E	3	204	200	-0,020	2.737	2.689	-0,018	2.041.101	1.737.920	-0,149
E	4	290	308	0,062	11.945	13.194	0,105	9.594.925	9.626.133	0,003
E	5	68	72	0,059	10.383	10.975	0,057	7.541.590	7.308.129	-0,031
E	6	60	62	0,033	110.144	115.409	0,048	40.874.412	41.892.417	0,025
E	*tot 1-6*	*1.202*	*1.202*	*0,000*	*137.491*	*144.513*	*0,051*	*62.611.370*	*62.611.370*	*0,000*
F	1	103.858	99.499	-0,042	234.242	231.743	-0,011	435.112.635	356.423.603	-0,181
F	2	18.697	19.330	0,034	134.492	139.260	0,035	245.290.296	216.466.307	-0,118
F	3	12.009	13.609	0,133	156.698	179.872	0,148	286.328.954	287.229.430	0,003
F	4	4.909	6.959	0,418	167.277	237.676	0,421	327.440.039	413.843.798	0,264
F	5	218	284	0,303	31.801	40.564	0,276	67.777.052	75.078.240	0,108
F	6	81	91	0,123	42.425	54.542	0,286	86.152.688	99.060.286	0,150
F	*tot 1-6*	*139.772*	*139.772*	*0,000*	*766.935*	*883.657*	*0,152*	*1.448.101.664*	*1.448.101.664*	*0,000*
G	1	130.940	124.846	-0,047	300.085	300.604	0,002	419.323.667	334.278.125	-0,203
G	2	23.062	24.127	0,046	165.591	174.297	0,053	169.891.137	151.814.874	-0,106
G	3	15.333	17.048	0,112	200.176	223.516	0,117	192.359.753	195.438.192	0,016
G	4	7.275	10.444	0,436	256.962	380.285	0,480	236.238.587	318.409.449	0,348
G	5	536	668	0,246	80.196	98.512	0,228	74.340.104	85.327.282	0,148
G	6	244	257	0,053	228.908	244.647	0,069	241.633.741	248.519.067	0,028
G	*tot 1-6*	*177.390*	*177.390*	*0,000*	*1.231.918*	*1.421.861*	*0,154*	*1.333.786.989*	*1.333.786.989*	*0,000*
H	1	41.479	39.530	-0,047	97.429	97.567	0,001	124.999.239	100.151.827	-0,199
H	2	7.835	8.122	0,037	56.103	58.481	0,042	55.721.800	48.481.273	-0,130
H	3	4.896	5.508	0,125	63.903	72.158	0,129	63.361.007	64.463.559	0,017
H	4	2.253	3.263	0,448	80.755	119.044	0,474	84.144.988	108.511.281	0,290
H	5	146	180	0,233	22.751	27.324	0,201	21.826.146	25.723.940	0,179
H	6	78	84	0,077	81.697	85.575	0,047	105.625.641	108.346.941	0,026
H	*tot 1-6*	*56.687*	*56.687*	*0,000*	*402.638*	*460.149*	*0,143*	*455.678.821*	*455.678.821*	*0,000*
I	1	22.762	21.827	-0,041	54.559	54.493	-0,001	83.145.316	68.956.427	-0,171
I	2	5.803	5.845	0,007	42.183	42.570	0,009	53.303.840	45.470.257	-0,147
I	3	4.742	4.864	0,026	63.353	65.245	0,030	80.083.095	70.988.373	-0,114
I	4	3.390	4.051	0,195	130.345	159.924	0,227	147.117.320	160.022.338	0,088
I	5	480	548	0,142	72.367	81.732	0,129	79.221.051	78.428.416	-0,010
I	6	254	296	0,165	294.217	327.878	0,114	344.870.097	363.874.908	0,055
I	*tot 1-6*	*37.431*	*37.431*	*0,000*	*657.024*	*731.842*	*0,114*	*787.740.719*	*787.740.719*	*0,000*

segue

Nota: $i = 0$ definisce la distribuzione originaria, all'iterazione nulla; $i = i^*$ definisce la distribuzione che si ottiene a terminazione del processo di selezione.

Tab. 7.2 segue

ATECO	cl. dim.	N i = 0	N i = i*	Variaz. N	Dip. i = 0	Dip. i = i*	Variaz. Dip	Spesa per i = 0	Spesa per i = i*	Variaz. spesa
J	1	6.243	5.936	-0,049	14.276	14.180	-0,007	24.397.309	19.560.238	-0,198
J	2	831	952	0,146	5.953	6.887	0,157	6.099.979	6.562.179	0,076
J	3	672	744	0,107	8.999	9.902	0,100	8.071.853	8.401.872	0,041
J	4	845	920	0,089	38.628	42.848	0,109	27.210.344	26.737.444	-0,017
J	5	234	258	0,103	35.931	39.988	0,113	22.358.528	23.009.484	0,029
J	6	244	259	0,061	364.573	386.230	0,059	141.546.277	145.413.073	0,027
J	*tot 1-6*	*9.069*	*9.069*	*0,000*	*468.360*	*500.035*	*0,068*	*229.684.290*	*229.684.290*	*0,000*
K	1	81.715	78.765	-0,036	158.344	159.264	0,006	251.176.441	204.440.489	-0,186
K	2	10.478	11.472	0,095	75.379	83.228	0,104	73.830.187	73.555.525	-0,004
K	3	7.818	8.319	0,064	103.343	109.338	0,058	98.847.835	95.372.107	-0,035
K	4	5.756	6.948	0,207	230.367	283.851	0,232	232.558.415	239.357.822	0,029
K	5	830	992	0,195	124.150	149.562	0,205	126.301.108	126.835.391	0,004
K	6	376	477	0,269	274.685	350.831	0,277	213.072.769	256.225.421	0,203
K	*tot 1-6*	*106.973*	*106.973*	*0,000*	*966.268*	*1.136.074*	*0,176*	*995.786.755*	*995.786.755*	*0,000*
L	1	135	137	0,015	388	351	-0,095	379.979	270.055	-0,289
L	2	104	93	-0,106	777	689	-0,113	594.745	366.781	-0,383
L	3	137	130	-0,051	1.853	1.771	-0,044	1.575.688	1.286.931	-0,183
L	4	161	172	0,068	6.919	7.717	0,115	5.812.520	5.398.154	-0,071
L	5	35	36	0,029	4.865	5.331	0,096	2.806.009	3.015.661	0,075
L	6	8	12	0,500	7.913	9.243	0,168	3.162.790	3.994.149	0,263
L	*tot 1-6*	*580*	*580*	*0,000*	*22.715*	*25.102*	*0,105*	*14.331.731*	*14.331.731*	*0,000*
M	1	4.591	4.366	-0,049	12.797	12.284	-0,040	13.607.380	10.394.351	-0,236
M	2	1.696	1.621	-0,044	12.232	11.732	-0,041	9.885.248	7.412.476	-0,250
M	3	1.233	1.254	0,017	16.603	17.023	0,025	12.952.438	11.316.546	-0,126
M	4	847	1.104	0,303	30.154	41.342	0,371	23.277.064	28.340.339	0,218
M	5	43	63	0,465	6.000	8.713	0,452	5.415.527	7.059.285	0,304
M	6	9	11	0,222	3.757	4.430	0,179	3.352.001	3.966.661	0,183
M	*tot 1-6*	*8.419*	*8.419*	*0,000*	*81.543*	*95.524*	*0,171*	*68.489.658*	*68.489.658*	*0,000*
N	1	16.905	16.115	-0,047	35.188	34.906	-0,008	49.884.582	38.589.161	-0,226
N	2	2.217	2.377	0,072	15.928	17.216	0,081	16.759.722	15.347.821	-0,084
N	3	1.869	2.153	0,152	25.172	28.833	0,145	28.202.856	28.679.406	0,017
N	4	2.157	2.469	0,145	91.424	104.499	0,143	111.484.069	116.940.417	0,049
N	5	373	408	0,094	55.483	60.962	0,099	79.849.137	81.046.426	0,015
N	6	132	131	-0,008	76.154	85.472	0,122	110.925.243	116.502.378	0,050
N	*tot 1-6*	*23.653*	*23.653*	*0,000*	*299.349*	*331.888*	*0,109*	*397.105.609*	*397.105.609*	*0,000*
O	1	24.345	23.190	-0,047	52.027	51.768	-0,005	67.421.238	52.109.490	-0,227
O	2	3.534	3.793	0,073	25.461	27.493	0,080	23.642.850	21.607.203	-0,086
O	3	2.495	2.849	0,142	33.320	37.757	0,133	32.376.587	31.967.976	-0,013
O	4	1.756	2.199	0,252	67.525	85.502	0,266	64.292.416	66.794.327	0,039
O	5	190	252	0,326	28.566	38.195	0,337	32.876.894	34.148.092	0,039
O	6	88	125	0,420	67.901	91.622	0,349	66.425.642	80.408.539	0,211
O	*tot 1-6*	*32.408*	*32.408*	*0,000*	*274.800*	*332.337*	*0,209*	*287.035.627*	*287.035.627*	*0,000*
Q	1	576	578	0,003	708	776	0,096	1.700.298	1.663.869	-0,021
Q	2	33	30	-0,091	242	217	-0,103	176.248	172.866	-0,019
Q	3	19	17	-0,105	252	225	-0,107	163.211	121.866	-0,253
Q	4	16	18	0,125	590	661	0,120	380.301	309.594	-0,186
Q	5	-	-	0,000	-	-	0,000	-	-	0,000
Q	6	-	1	-	-	379	-	-	151.863	-
Q	*tot 1-6*	*644*	*644*	*0,000*	*1.792*	*2.258*	*0,260*	*2.420.058*	*2.420.058*	*0,000*
NR	1	374	363	-0,029	876	861	-0,017	1.237.375	1.011.617	-0,182
NR	2	77	81	0,052	549	576	0,049	853.735	556.464	-0,348
NR	3	66	66	0,000	863	873	0,012	973.704	858.878	-0,118
NR	4	50	54	0,080	1.884	2.159	0,146	1.784.703	1.837.505	0,030
NR	5	2	4	1,000	294	575	0,956	273.263	500.895	0,833
NR	6	2	3	0,500	577	974	0,688	502.644	860.065	0,711
NR	*tot 1-6*	*571*	*571*	*0,000*	*5.043*	*6.018*	*0,193*	*5.625.424*	*5.625.424*	*0,000*

Nota: $i = 0$ definisce la distribuzione originaria, all'iterazione nulla; $i = i^*$ definisce la distribuzione che si ottiene a terminazione del processo di selezione.

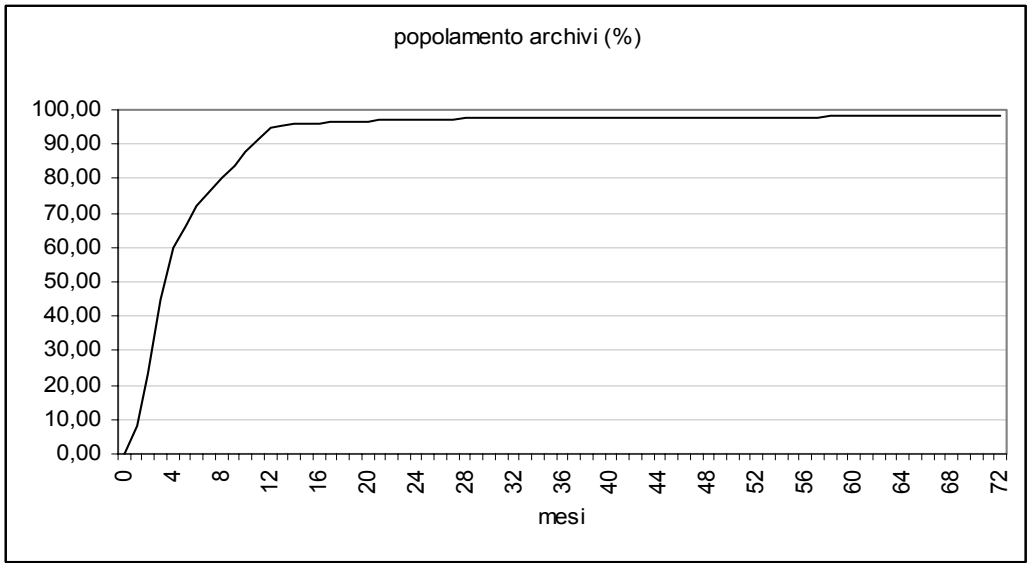
8. Il popolamento dei DM10, tempi di attesa, metodo di donazione

Prima della lettura di questo paragrafo si consideri che la soluzione al problema non è ancora disponibile. Quanto segue è quindi frutto di analisi preliminari ed è coerente con una linea di attività ancora in corso di svolgimento. Si inseriscono comunque in questa occasione le linee generali di approccio.

Gli archivi dell'universo DM10 sono soggetti ad un popolamento progressivo, funzione della distanza temporale dalla data di riferimento. Ciò è dovuto ai ritardi di trasmissione dell'informazione cartacea da azienda a sede INPS di

competenza e da questa ultima alla sede centrale, che traduce l'informazione su supporto informatico.
Allo stato attuale le imprese che utilizzano ancora la dichiarazione cartacea costituiscono circa i due terzi del totale.
Assumendo come universo delle imprese esistenti all'anno t quelle presenti negli archivi in $t+72$ mesi (sei anni dopo l'anno di riferimento), è stato verificato che la dinamica del popolamento degli archivi rispetto al tempo segue una traiettoria logistica con raggiungimento del 100% a 72 mesi dall'anno di riferimento (il raggiungimento del 100% è implicito alla definizione di totale).
Il grafico 7.1 riproduce l'andamento temporale del popolamento.

Figura 8.1. Andamento nel tempo del popolamento degli archivi DM10



Come si nota, l'informazione a 13 mesi dal periodo di riferimento garantisce una copertura di oltre il 95% del totale calcolato a 72 mesi. L'andamento temporale del popolamento oltre i 13 mesi presenta un dinamica piuttosto piatta, il che significa che l'estensione del periodo di attesa dei dati produce guadagni trascurabili nella rappresentazione statistica del fenomeno. Una estensione del periodo di acquisizione degli archivi a due anni si associa ad un guadagno di soli 3 punti percentuali (da circa il 95% a circa il 98%).
Per gli stessi motivi richiamati nel paragrafo precedente, si ritiene che il guadagno informativo ottenuto con l'estensione dei tempi di popolamento a più di 13 mesi (tempo dopo il quale la curva assume un andamento piatto), in considerazione delle perdite di contenuto informativo connesse al necessario ritardo di produzione, non sia giustificato. Si opta pertanto per la fissazione dei tempi di attesa a 13 mesi dal periodo di riferimento.
L'individuazione del rimanente 5% procede quindi attraverso tecniche di *donazione*.
Poiché l'insieme del 5% mancante è costituito principalmente da aziende di piccole dimensioni, (che utilizzano con maggiori difficoltà i supporti informatici), per esse è ragionevole ritenere che sia presente un elevato grado di "turbolenza", cioè un forte tasso di natalità e mortalità sul totale. La tabella 8.1 mostra i tassi di natalità e mortalità osservati nella transizione dall'archivio

DM10 relativo all'anno 2000 a quello relativo all'anno 2002, distinti per classi dimensionali.

Tab. 8.1: Natalità e mortalità delle unità di analisi (matricole), per classi di dipendenti

Classe dim.	Totale matricole	Matricole entranti	%	Matricole uscenti	%
1	1.181.855	246.277	20,84	163.208	13,81
2	114.610	3.583	3,13	3.963	3,46
3	93.139	2.306	2,48	2.592	2,78
4	60.440	1.394	2,31	1.633	2,70
5	6.935	182	2,62	209	3,01
6	3.114	94	3,02	110	3,53
tot	1.460.093	253.836	17,38	171.715	11,76

Nota: transizione anni 2000-2002, differenze insiemistiche

Come si nota, la turbolenza si addensa nella prima classe dimensionale.
Questa considerazione limita la possibilità di utilizzare, nella ripetizione annuale delle procedure, l'informazione sul 5% deducibile da una estensione dei tempi di popolamento effettuata su anni di riferimento precedenti a quello di analisi. Il rischio è quello dell'attribuzione di valori ad imprese di fatto non più attive.
Si opta pertanto per un metodo di donazione "puro", basato esclusivamente sull'informazione sezionale e non su quella temporale.
A tal fine, l'informazione disponibile per il 95% viene affiancata da informazioni dicotomiche rispetto ai caratteri regione e settore ATECO e da informazioni continue rispetto ad occupazione e fatturato, provenienti da fonti ASIA e DM10. In una prima fase esse costituiscono, per ogni trattamento, le variabili esplicative della probabilità di effettuare almeno una erogazione nell'anno.
Il contributo marginale dei fattori esplicativi alla definizione di tale probabilità (la variabile trattamento assume valore dicotomico, 1 se viene effettuato almeno un trattamento, 0 altrimenti) viene stimata ricorrendo a tecniche *QR* (qualitative response) del tipo *logit* o *probit*[22]. La funzione di probabilità stimata ha la forma generica:

$$P(y_i = 1) = \sum_{r=1}^{20} \alpha_r D_{i,r} + \sum_{s=1}^{17} \beta_s S_{i,s} + \gamma N_i + \delta F_i + \varepsilon_i$$

dove *S* è una matrice contenente le *dummies* di settore, mentre *N* ed *S* (variabili continue) contengono, rispettivamente, osservazioni su occupazione alle dipendenze e fatturato aziendale.
In una seconda fase, dal confronto delle imprese del 95% con l'archivio delle imprese attive in ASIA, si definisce, per differenza, l'insieme delle imprese *potenzialmente* esposte al rischio di appartenere al rimanente 5%, cioè di non essere presenti nell'archivio a 13 mesi, pur avendo effettuato pagamenti per TMNP.
Tra queste, ne viene estratta una numerosità esattamente corrispondente al 5% del totale potenziale, attraverso l'utilizzo di tecniche di campionamento. È auspicabile che gli strati del campionamento siano quelli dall'informazione proveniente dall'analisi dei contributi marginali prodotti dall'analisi *QR*.

[22] Per una rassegna dei metodi si veda Ameniya (1981) e Maddala (1983).

Estratto il campione, ad ognuna di queste imprese - per ogni trattamento - viene attribuito il valore medio di erogazione dedotto dalle imprese aventi caratteristiche analoghe presenti nel 95%.
Terminata questa ultima fase di correzione si ottiene pertanto l'archivio consolidato dei TMNP.

9. Considerazioni tecnologiche

La sperimentazione finora descritta ha permesso non solo di definire e realizzare un nuovo approccio statistico-metodologico sugli archivi INPS-DM10, ma anche di valutare l'impatto di alcune scelte tecnologiche.
Gli archivi DM10, nella fase di acquisizione, si presentano come data-set SAS mensili di medie dimensioni (2-3 *Gb*/mese). La dimensione complessiva dei file annuali (25 *Gb*/anno) e la complessità delle procedure di elaborazione e consolidamento descritte nel documento hanno portato ad un ripensamento complessivo dell'architettura informatica.
La sperimentazione è stata condotta secondo un approccio *try-and-repeat* (ogni ipotesi formulata deve essere sottoposta a verifica ed eventualmente modificata in base ai risultati ottenuti) e ciò implica che l'intero processo di elaborazione deve poter fornire risposte in tempi contenuti lavorando sull'universo dei dati. Queste necessità hanno imposto l'adozione di un approccio RAD (*Rapid Application Development*) che ha determinato precise scelte architetturali:

→ gli spazi di lavoro necessari (150*Gb*) hanno escluso l'utilizzo sia degli attuali server di Istituto, per la loro cronica mancanza di spazio, sia l'ambiente PC/Windows non pensato per archivi di queste dimensioni.
→ l'ambiente di lavoro da utilizzare (linguaggi di programmazione) deve poter offrire una ragionevole flessibilità nella lavorazione ed incrocio degli archivi, e ciò ha fortemente limitato l'utilizzo dell'ambiente SAS.
→ L'acceso a librerie di elaborazione statistico-scientifica non soggette a licenze chiuse (a pagamento), ha portato all'utilizzo di ambienti OSS (*Open Source Software*).

L'architettura risultante è costituita da ambienti integrati, dove ognuno dei quali non offre risposte a tutte le esigenze, ma risulta specializzato solo in alcuni ambiti elaborativi. Sfruttando la possibilità di accedere all'interno dell'Istituto ad una macchina Linux Red-Hat su architettura Intel, è stato configurato il seguente ambiente:

→ *Ambiente hardware*: architettura Intel, biprocessore, 3*Gb* di Ram, 2 dischi da 160*Gb*.
→ *Sistema operativo*: Red Hat server edition 7 (www.redhat.com).
→ *DBMS*: Oracle 9 (www.oracle.com).
→ *Ambiente di sviluppo*: GCC 3.3 (gcc.gnu.org), strumenti GNU correlati (www.gnu.org/directory/GNU).
→ *Librerie di calcolo scientifico*: GSL – Gnu Scientific Library (www.gnu.org/software/gsl).

Questa architettura presenta numerosi vantaggi:

→ il progetto TMNP condivide agevolmente le risorse disponibili con altri progetti di dimensioni equivalenti o superiori.
→ ognuna delle componenti descritte può essere sostituita con un'altra equivalente, e ciò non compromette il resto del sistema.
→ ogni postazione di lavoro può collegarsi alla banca dati finale utilizzando gli strumenti più largamente utilizzati in Istituto (SAS, Toad, ..).
→ il costo dell'intera architettura è molto vicino a quello di 2-3 PC.

È necessario sottolineare come non esistano soluzioni assolute ed anche questo approccio ha dei limiti ed inconvenienti, ma essi sono principalmente legati al background culturale degli attori del processo.

10. Linee di sviluppo futuro

L'incrocio degli archivi mensili DM10, degli archivi annuali del servizio OCC e del servizio ARC ha permesso la costruzione di un archivio consolidato basato sulle seguenti variabili: anno di riferimento, sezione ateco, macrosettore, classe dimensionale, funzione, prestazione, trattamento, forma di pagamento, importo.
Possibili estensioni del contenuto informativo di questo archivio potranno riguardare i seguenti aspetti:

→ *forma giuridica*: tale variabile esiste nella sezione anagrafica dell'archivio mensile DM10, ma necessita di una robusta procedura di normalizzazione e classificazione delle voci in essa presente;
→ *anagrafica di impresa*: ai fini di una migliore valutazione della dinamica delle imprese (nascite, cessazioni, scissioni, fusioni) è necessario introdurre informazioni che permettano di seguire l'intero ciclo di vita di una azienda intesa come posizione contributiva INPS (matricola). Questa anagrafica permetterebbe la considerazione di un dettaglio mensile nelle analisi (attualmente è annuale);
→ *regionalizzazione del dato*: il fenomeno dell'"*accentramento contributivo*"[23] rende difficile la ricostruzione della distribuzione territoriale della spesa TMNP. Il problema è risolvibile costruendo, per le imprese autorizzate all'accentramento, una matrice di pesi con le seguenti caratteristiche:
    - le righe rappresentano le matricole INPS;
    - le colonne rappresentano le province italiane;
    - le celle di incrocio riportano il numero di dipendenti che lavorano nella provincia/matricola considerata.

11. Conclusioni

Il documento ha voluto fornire una descrizione sintetica e non eccessivamente tecnica del processo di elaborazione dell'informazione sui trattamenti monetari di tipo non pensionistico erogati dalle imprese per conto dell'INPS. Si è dato maggiore rilievo ad una descrizione puntuale delle ipotesi e del significato delle tecniche implementate ai fini del controllo e correzione del dato. Nella costruzione delle procedure si è privilegiato l'automatismo degli interventi, implementando parallelamente una serie di indicatori degli effetti degli stessi che ne permettono il controllo e la trasparenza delle operazioni. L'applicazione

[23] Le aziende plurilocalizzate possono richiedere di centralizzare la dichiarazione della spesa TMNP verso un solo ufficio provinciale INPS.

dell'insieme delle procedure ai dati non genera variazioni notevoli di spesa. I maggiori effetti vengono prodotti piuttosto sul popolamento delle diverse classi dimensionali. Ciò è dovuto all'algoritmo di identificazione e correzione del dato occupazionale, di fonte esterna all'archivio, della cui opportunità si è data giustificazione. Nell'insieme, l'immagine che si ottiene risulta più vicina alla distribuzione per classi desumibile dagli archivi consolidati sull'universo delle imprese italiane. Si consideri tuttavia che quelli forniti sono i risultati di esercitazioni sperimentali, soggette a modifica nelle applicazioni a regime. La forte parametrizzazione delle relazioni degli algoritmi di calcolo permette infatti la calibrazione dell'intervento in relazione alle diverse considerazioni di opportunità tecnica e teorica che si presenteranno nel corso delle indagini.

Riferimenti bibliografici

- Ameniya, T., 1981, "Qualitative Response Models: A Survey", *Journal of Economic Literature*, n.19.

- Baldi, C., Cimino, E., Pallata, A., Succi, R., Tuzi, D., 2001, "Una sperimentazione dell'utilizzo delle dichiarazioni mensili e dell'anagrafe delle posizioni contributive dell'INPS per la stima trimestrale dell'occupazione dipendente", documento di lavoro INPS A05 del progetto interarea*: utilizzo dei dati INPS per la produzione di statistiche correnti su occupazione e retribuzioni, Istat.*

- Barnett, V. e Lewis, T., 1994, *Outliers in Statistical Data*, third edition, Wiley & Sons, New York.

- Cimino, E., Succi, R., Tuzi, D., 2000, "La ricostruzione delle imprese a partire dalle posizioni contributive degli archivi INPS", documento di lavoro INPS C07 del progetto interarea*: utilizzo dei dati INPS per la produzione di statistiche correnti su occupazione e retribuzioni, Istat.*

- Consolini, P., De Carli, R., 2002, "Le prestazioni sociali monetarie non pensionistiche: unità di analisi, fonti e rappresentazione statistica dei dati", *Documenti Istat*, Istat, Roma.

- Douglas Faires, J. and Burden, R., 1998, *Numerical Methods*, Brooks/Cole Publishers, UK.

- Eurostat, 1996, Esspros Manual, Population and social conditions, Methods, Luxemburg.

- Falorsi, P.D., Pallara, A e Russo, A. (a cura di), 2003, *Temi di ricerca ed esperienze sull'utilizzo a fini statistici di dati di fonte amministrativa*, Franco Angeli, Roma.

- Gismondi, R., 2002, "Confronti tra metodi per l'individuazione di osservazioni anomale in indagini longitudinali: proposte teoriche e verifiche empiriche" in *Quaderni di Ricerca*, 1, Istat, Franco Angeli, Roma.

- Hidiroglou, M.A. e Berthelot, J.M., 1986, "Statistical Editing and Imputation for Periodic Business Surveys", in *Survey Methodology*, 12, Statistics Canada, Ottawa.

- Istat, 1991, *Classificazione delle attività economiche*, Metodi e norme, serie C – n. 11, Istat, Roma.

- Maddala, G.S., 1983, *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.
- www.gnu.org/directory/GNU, Ambiente di sviluppo open source GCC.
- www.gnu.org/software/gsl, Librerie di calcolo scientifico: GSL – Gnu Scientific Library.
- www.oracle.com, DBMS: Oracle 9.
- www.redhat.com, Red Hat server edition 7.